How do NCLB's allowances for state discretion affect AYP determinations? To answer this question, we start at the end of the story, by first reporting how our sample of schools performed in the various states relative to making AYP. Next, we explain the components that contributed to this judgment.

How the Sample Performed Relative to State AYP Requirements

Table 1 summarizes the performance of our elementary and middle school samples in making AYP in 2008 across the 28 states we studied. With 18 elementary and 18 middle schools, there were 504 opportunities to make or not make AYP at the elementary level (18 schools x 28 states) and 468 opportunities at the middle school level (18 schools x 26 states).

Table 1. Proportion of schools in the sample that met AYP requirements in 2008

School type	Number and percentage of schools making AYP
Elementary schools	159/504 (32%)
Middle schools	52/468 (11%)

The table shows that our elementary schools made AYP less than one-third of the time. But our middle schools did even worse, making AYP in just over one in ten cases.

Within the elementary school sample, the number of schools that made AYP varied greatly by state. In Massachusetts and Nevada, only one school made AYP, while in Wisconsin, 17 of the 18 schools did (Figure 1). To rephrase, in Massachusetts and Nevada, almost none of

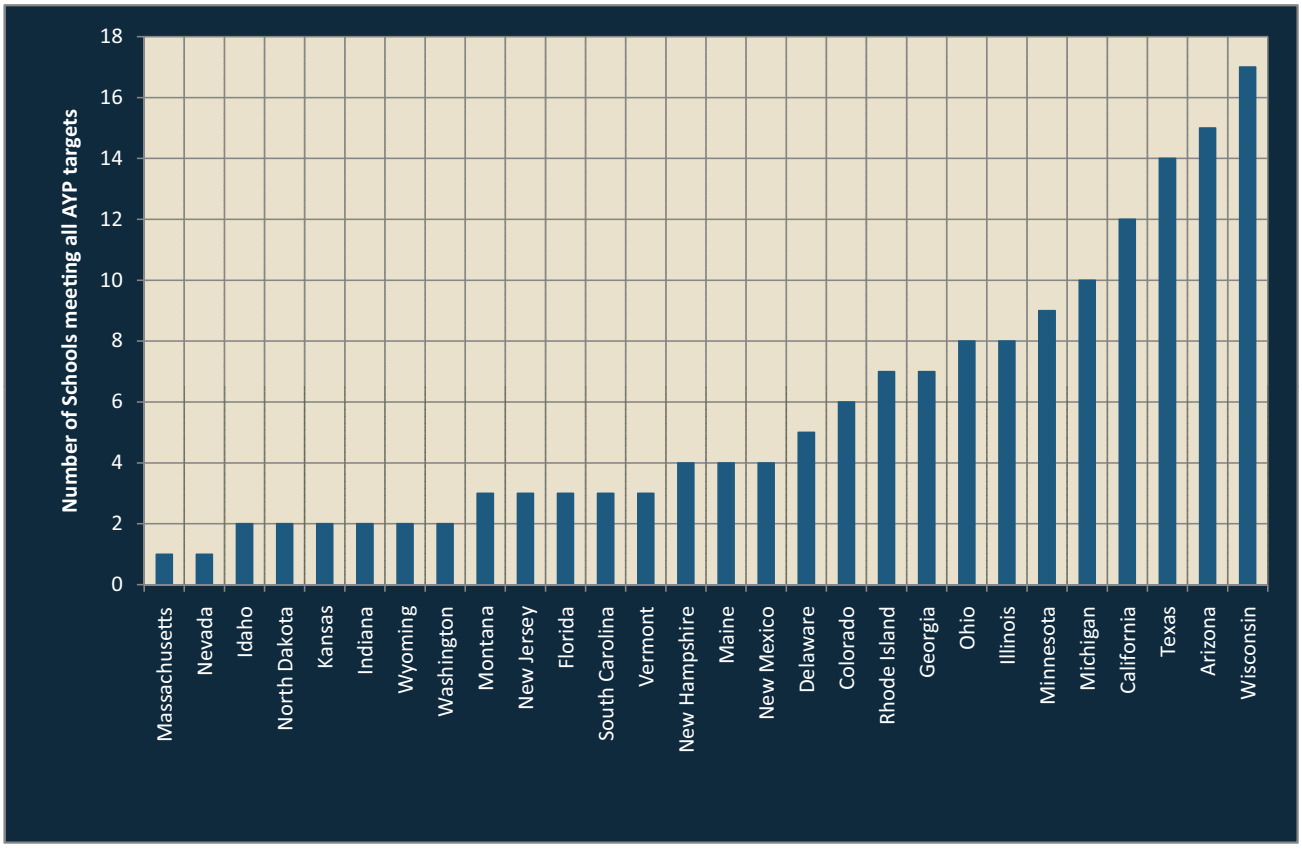

Figure 1. Number of schools in the elementary school sample making AYP by state (2008)

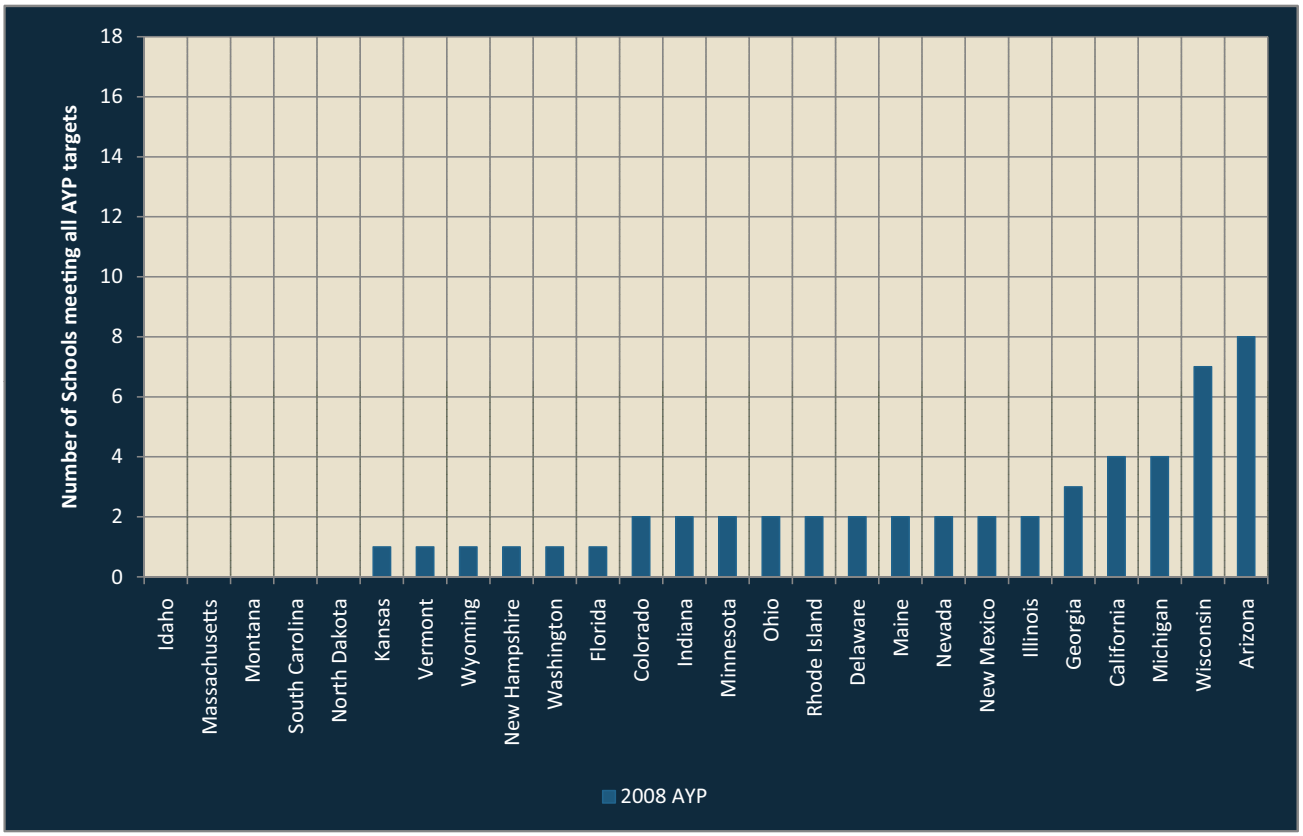

Figure 2. Number of schools in the middle school sample making AYP by state (2008)

Note: Texas and New Jersey are not included in the middle school analysis since cut score estimates for 8th grade were not available in these states.

the elementary schools in our sample made AYP, while in Wisconsin, almost all of them did. **Keep in mind that these are the exact same schools**.

There was more consistency across states with the middle school sample because the vast majority of schools failed to make AYP in most of the states (see Figure 2). In 21 of the 26 states we studied, two or fewer schools met the 2008 AYP requirements. In no state did half of the middle schools meet the 2008 AYP requirements.

The disappointing performance of the schools in the sample led to the questions that ultimately drove the study. For the elementary school sample, why were the AYP outcomes for the group so different across states? For the middle school sample, why did so many fail to make AYP?

The answers to these questions are found in an analysis of three factors that affect whether schools make AYP. These are:

1. The interaction between proficiency cut scores in math and reading and the difficulty of the AMOs;

2. The application of a confidence interval (i.e., margin of error); and

3. The performance of various subgroups, and whether they count for accountability purposes. These subgroups include low-income students, traditionally disadvantaged minorities, limited English proficient (LEP) students, and students with disabilities (SWDs).

In the following subsections, we discuss each of these factors in turn.

The Interactions between Cut Scores and AMO Difficulty (Factor 1, Part 1)

The likelihood that a school will meet an annual target

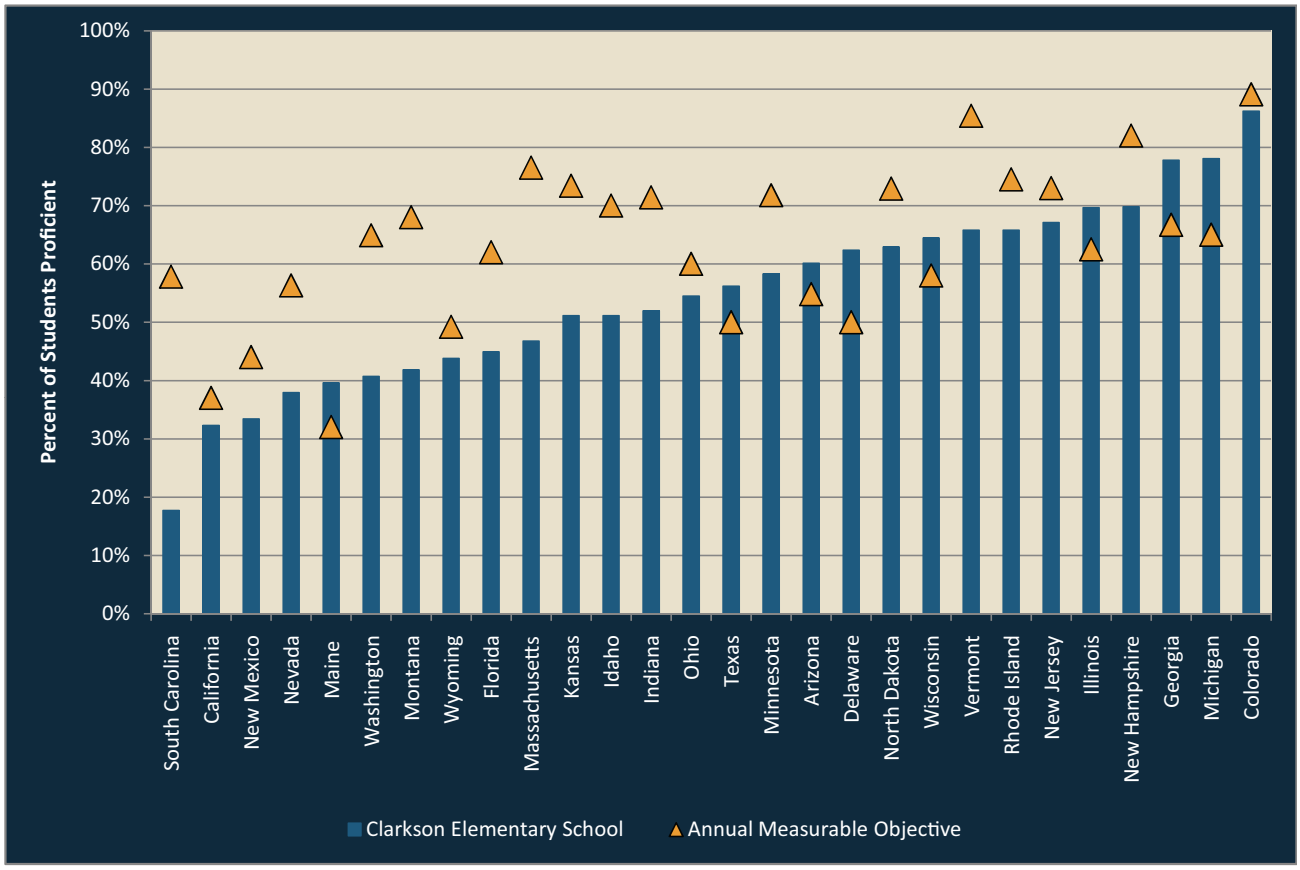

Figure 3. Math proficiency rate of Clarkson students relative to 2008 AMOs

Note: The length of the blue bar represents the percentage of Clarkson students who would be considered proficient in each state. The orange triangle represents the Annual Measurable Objective, or percentage of students required to be proficient in 2008 for the school to make AYP.

is strongly affected by two variables. The first is the difficulty of the test itself. In this case, we aren't talking about the content of the test (which is outside the scope of this study) but instead how difficult or easy it is for students to reach its passing score. The AMOs (i.e., the proportion of students in the school—and in each of the school's subgroups—that must pass the test each year) make up the second variable.

You can have an easy test and a difficult objective. For example, requiring a golfer to make a two-foot putt would be an easy proficiency test in that sport, but asking the same golfer to make 100 two-foot putts in a row would be a difficult objective.

The Case of Clarkson Elementary - Inconsistent proficiency rates and annual targets send conflicting signals

To illustrate this interaction, consider the case of one of our sample schools, Clarkson Elementary, a very diverse school serving primarily low-income students. Ninety-five percent of Clarkson students come from traditionally disadvantaged minority groups (African American, American Indian, and Hispanic/Latino), and 87% qualify for the low-income subgroup. Clarkson is the lowest performing elementary school in the sample. When compared to the NWEA norm group—a sample of over 1.2 million students who attend schools in 32 states (NWEA 2005)—Clarkson students perform, on average, 9.4 scale score points below the norm group's median in math and reading. This would mean that a typical sixth grader at Clarkson performs midway between the fourth grade and fifth grade NWEA norm median in these subjects. In our study, fall to spring scale score growth among Clarkson students was the lowest among the sampled elementary schools; its students attained only 55% of the average growth of students who started with equivalent scores on the NWEA assess-

ments. Setting aside the question of whether Clarkson elementary is a good or a bad school, we would nonetheless expect accountability metrics to identify Clarkson as a school in need of help.

Figure 3 shows the percentage of Clarkson's students who would be projected to reach the proficient level in math (indicated by blue bars) relative to the 2008 AMOs (indicated by the orange triangles) for the states we studied. Clarkson's projected math proficiency rate varied from 18% in South Carolina to 86% in Colorado (which uses "partially proficient" as its standard for NCLB proficiency). Clarkson's proficiency rate was sufficient to exceed the AMOs in 8 of the 28 states studied. So even though this was the lowest performing elementary school in our sample, Clarkson's performance in 2008 would still be considered adequate in eight states. More importantly, we can see very large differences in the percentage of Clarkson students who would be found proficient across states, and equally large differences in how AMOs are set.

In Clarkson's case, the differences in the math proficiency rates and AMOs conspire to send conflicting messages about student achievement based on the state in which the school is placed. If Clarkson were located in South Carolina, for example, its projected results on the state's current assessment (the Palmetto Achievement Challenge Tests, or PACT) would signal that the school's performance is entirely inadequate. Proficiency standards (i.e., the placement of cut scores) in South Carolina are challenging—only 18% of Clarkson students would have passed—and South Carolina's AMO requires 58% of students to pass. The resultant gap (Clarkson's pass rate would need to improve by 40 percentage points just to reach the AMO for 2008) would lead district administrators to conclude that major changes were needed. Overcoming such failure would likely require profound changes in the school's curriculum, culture, and staffing.

When we move Clarkson to Rhode Island, the situation looks far less bleak. Clarkson's math proficiency rate improves from 18% to 67%, a level of performance that fell within a stone's throw of the school's AMO (73%). We can envision incremental improvements to address this kind of gap, perhaps a school improvement plan focused on students' primary deficits. Parents and others reviewing achievement at Clarkson might not believe that performance is that bad, and relatively modest changes might, at least temporarily, fix the school's ailing proficiency rate.

Now, let's move Clarkson to Michigan. Here, math achievement seems to be just fine. More than three-quarters of the students (78%) are projected to achieve proficiency, a level of performance that is well beyond the 2008 AMO (65%). In such a setting, math achievement of the student body as a whole would hardly be a problem, and Clarkson's efforts would be focused on particular subgroups, if any, that may have failed to meet their AMOs.

Unfortunately, things at Clarkson are not fine. Not only is student achievement low, but students are making less progress than their peers. The problem is not limited to small enclaves of minority students, LEP pupils, or students with disabilities either; low achievement persists in all of the school's subgroups. But the messages delivered via accountability systems are highly inconsistent for schools like Clarkson across the country. In some states, the school is on an inevitable path to closure or reconstitution. In others, the problems seem solvable with an educational tweak here or there, and in a few states, there appears to be no problem at all.

Interactions between Cut Scores and AMOs Across the States (Factor 1, Part 2)

As we explained earlier, a school's likelihood of making AYP is affected by the interaction between the proficiency cut scores and the AMOs. Now we examine how this interaction played out in the various states in our study.

Figure 4 illustrates the difficulty of the various state cut scores in math by showing how our sample of eighteen elementary schools performed relative to those targets. In the majority of the states studied, schools are evaluated according to the proportion of students who achieve proficient (or better) on the state test. These states are represented by blue bars in the figure. Six of

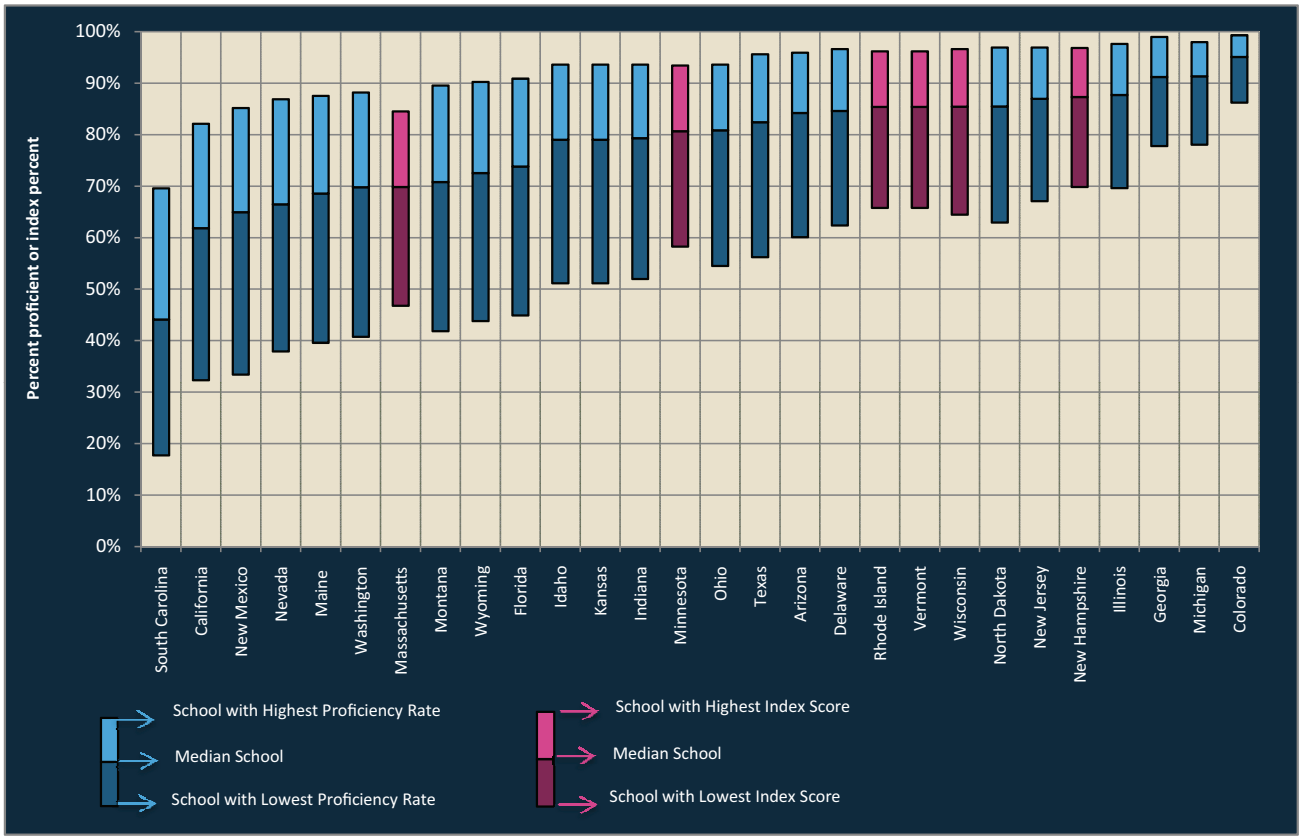

Figure 4. Overall proficiency rates of the elementary school sample in math

Note: Length and color of the bar represent the difference in overall performance between the highest and lowest performing school in each state. In South Carolina, for example, the lowest performing elementary school in the sample achieved an estimated proficiency rate of about 18% (as represented by the bottom of the dark blue section of the bar), the median school achieved 43% proficiency (marked by the line between the light and dark blue sections of the bar), and the highest performing school achieved 70% proficiency (shown by the top of the light blue section). States with higher cut scores are generally located at the left end and those with lower cut scores at the right. Magenta colored bars represent states that award students partial credit for achieving at lower proficiency levels.

the states studied (the magenta bars) use an index that gives full credit to students who achieve proficient (or better) and partial credit to students who perform at lower levels. The "index scores" in states using this hybrid model are always higher than the actual proficiency percentage.[1]

The length of the bar in Figure 4 represents the difference in overall performance between the lowest and highest performing sample school in the state. The middle line shows the performance of the median school in the sample. States are ordered by the performance of the median school; consequently, states with higher cut scores are generally located at the left end of the graph, and those with lower cut scores at the right. In South Carolina, for example, the lowest performing elementary school in the sample achieved an estimated proficiency rate of about 18% (as represented by the bottom of the dark blue section of the bar), the median school achieved 43% proficiency (marked by the line between the light and dark blue sections of the bar), and the highest performing school achieved 70% proficiency (shown by the top of the light blue section). By contrast, in Colorado, the lowest performing school achieved 88% proficiency, the median school achieved 95% proficiency rate, and the highest performing school achieved 99%.

[1] The six states studied that use an index are Rhode Island, Massachusetts, Minnesota, Vermont, Wisconsin, and New Hampshire. The index gives full credit to students who achieve proficient (or better) and partial credit to students performing at lower levels. Consequently, the resultant score in states using this "hybrid" model is always higher than the actual proficiency percentage (giving students partial credit for achieving lower proficiency levels is obviously better than no credit, at least for the schools' ratings). The index provides a fair amount of help when annual targets are below 50%; however, once targets rise above 75%, the index has far less impact.

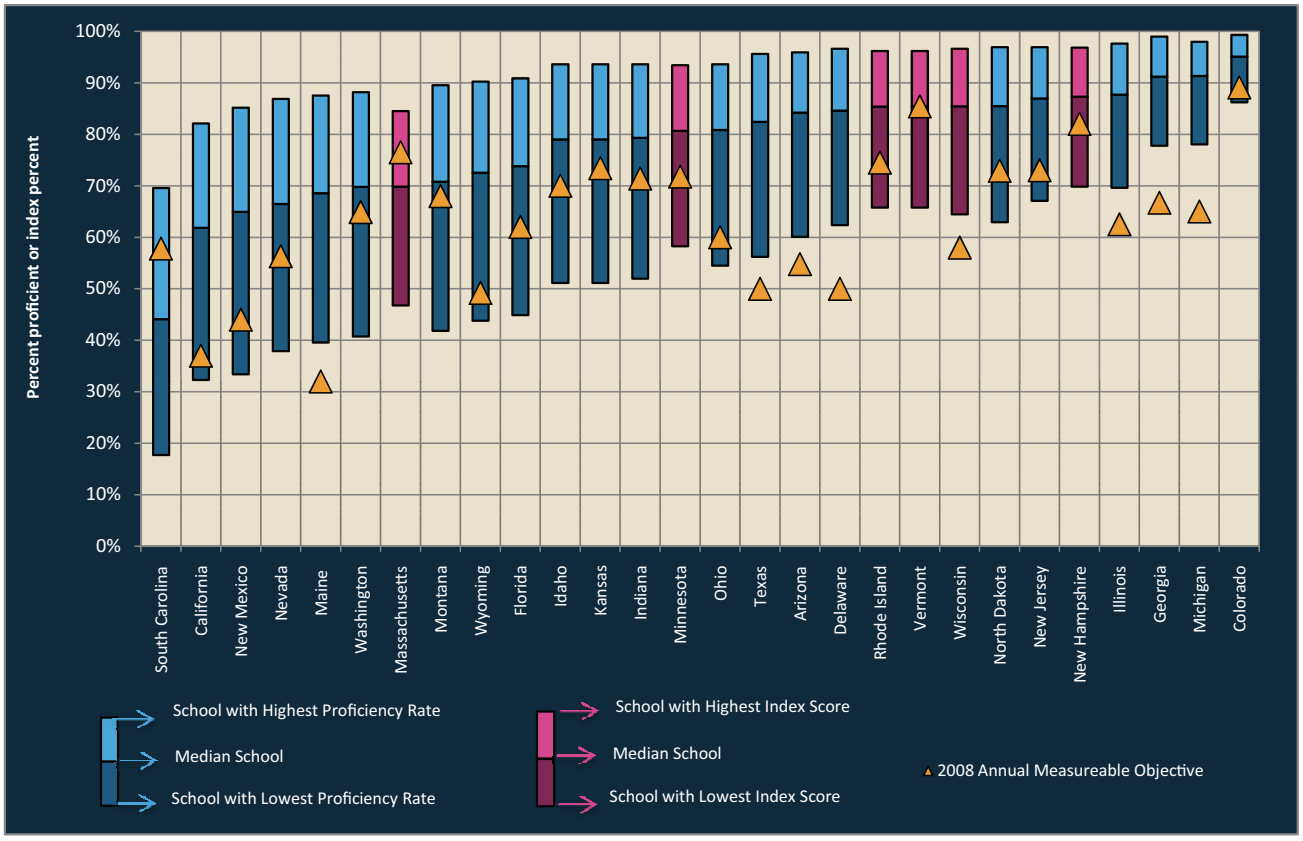

Figure 5. Math proficiency rates of the elementary school sample relative to each state's 2008 AMOs

Note: Length and color of the bar represent the difference in overall performance between the highest and lowest performing school in each state. In South Carolina, for example, the lowest performing elementary school in the sample achieved an estimated proficiency rate of about 18% (as represented by the bottom of the dark blue section of the bar), the median school achieved 43% proficiency (marked by the line between the light and dark blue sections of the bar), and the highest performing school achieved 70% proficiency (shown by the top of the light blue section). States with higher cut scores are generally located at the left end and those with lower cut scores at the right. The magenta bars represent states that award students partial credit for achieving at lower proficiency levels. The orange triangle represents the Annual Measurable Objective, or the proportion of students required to be proficient for the 2007–2008 school year. When the triangle is below the bar, all schools in the sample met that state's AMO.

Put another way, fewer than half the schools in our sample would have achieved a 50% proficiency rate if the schools were placed in South Carolina. Had these same schools been located in Georgia, Colorado, or Michigan, the top half of schools would all have achieved estimated proficiency rates greater than 90% (in each of those states, the line dividing the dark and light blue sections of the bar is above 90%).

It's no surprise that the proficiency rates varied from state to state in this study. This finding is consistent with any number of previous studies (McGlaughlin, et al. 2008; Cronin, et al 2007a; National Center for Educational Statistics 2007; Kingsbury, et al. 2003). But the cited studies reflect only one dimension of the assessment, the difficulty of the cut score. The difficulty of the AMOs must also be considered, as we've done in this research.

Figure 5 adds the 2008 AMOs (orange triangles), which show the percentage of students who must be proficient in order for the school to make AYP. The placement of the AMO triangles allows us to see the proportion of the sample that met its target. We can see, for example, that South Carolina's 2008 AMO requires a proficiency rate of 58%. About one-quarter of the sample schools achieved this rate of proficiency. This tells us that South Carolina's proficiency cut score is high relative to the other states and that its AMO is also quite challenging.

Our Michigan results showed the opposite case—Michigan's AMO requires a proficiency rate of 65%, but all

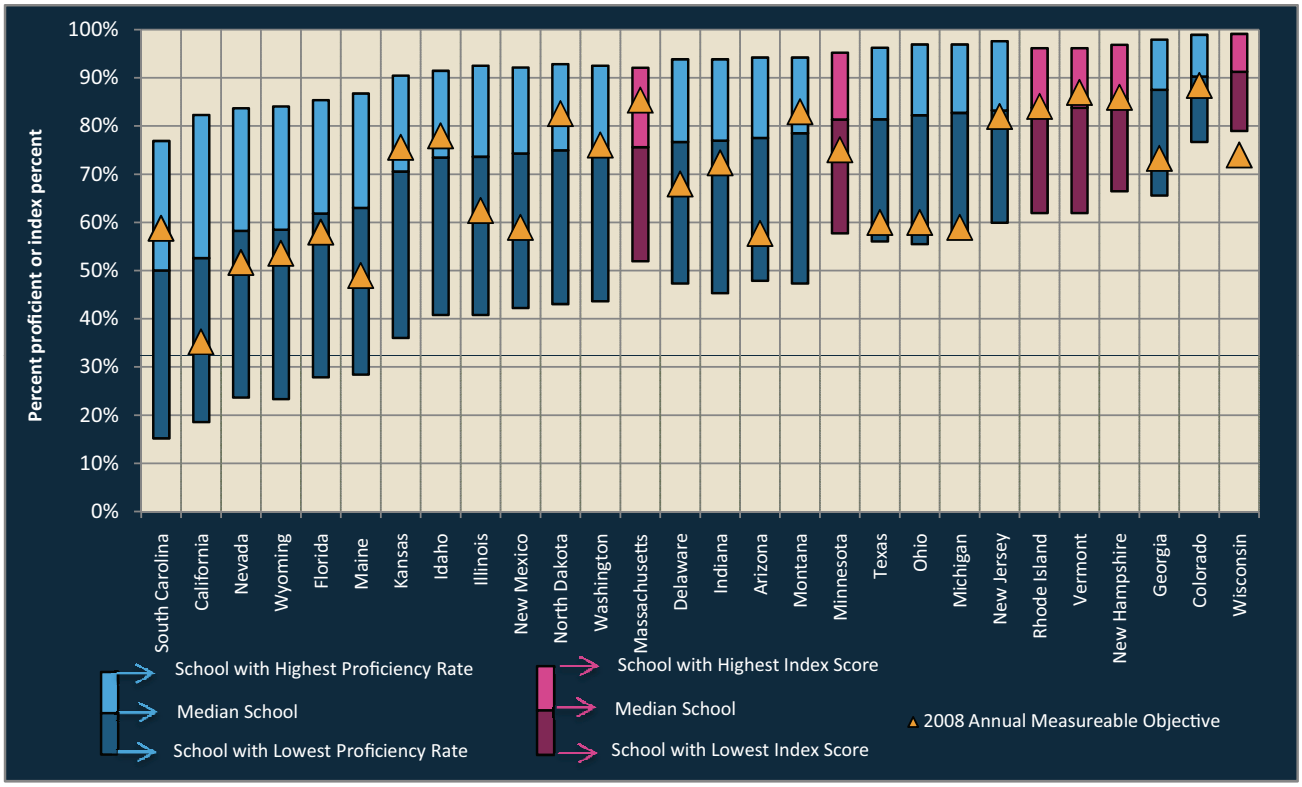

Figure 6. Reading proficiency rates of the elementary school sample relative to each state's 2008 AMOs

Note: Length and color of the bar represent the difference in overall performance between the highest and lowest performing school in each state. In South Carolina, for example, the lowest performing elementary school in the sample achieved an estimated proficiency rate of about 18% (as represented by the bottom of the dark blue section of the bar), the median school achieved 43% proficiency (marked by the line between the light and dark blue sections of the bar), and the highest performing school achieved 70% proficiency (shown by the top of the light blue section). States with higher cut scores are generally located at the left end and those with lower cut scores at the right. The magenta bars represent states that give students partial credit for achieving at lower proficiency levels. The orange triangle represents the Annual Measurable Objective, or the proportion of students required to be proficient for the 2007–2008 school year. When the triangle is below the bar, all schools in the sample met that state's AMO.

schools in the sample achieved well beyond this level (indicated by the blue bar floating above the AMO triangle). Keep in mind that we're referring here to schools as a whole reaching their AMOs; we haven't yet considered the impact of subgroup performance. Thus, not only is the Michigan cut score low relative to the other states (remember that states with lower cut scores generally appear on the right), but its AMO is low as well. We could contrast Michigan with Colorado, which reports higher proficiency rates than Michigan (primarily because Colorado gives credit for "partially proficient" students), but has a considerably higher AMO (compare the placement of the orange triangles).

Schools must meet AMOs in both math and reading, so Figure 6 shows the results for the elementary school sample in reading. In general, the AMOs for reading are higher than those for math in the elementary school sample. Although all schools met the math AMOs in eight states (see Figure 5), there was only one state, Wisconsin, in which the entire sample met the reading AMO (indicated by the magenta bar floating above the AMO triangle). In 8 of the 28 states, fewer than half of the schools achieved the AMOs.

Once again, states with relatively low cut scores do not always have easy AMOs. Colorado's AMO was achieved only by about half of the sample, while the AMOs for Wisconsin and Georgia—other states with low cut scores—were achieved by all (Wisconsin) or nearly all (Georgia) schools (note placement of the orange triangles in Figure 6).

Math and reading proficiency rates for the middle school sample were typically lower than those for elementary schools, but AMOs in the states are set at a level that mitigated some of these differences. In seven states (Ari-

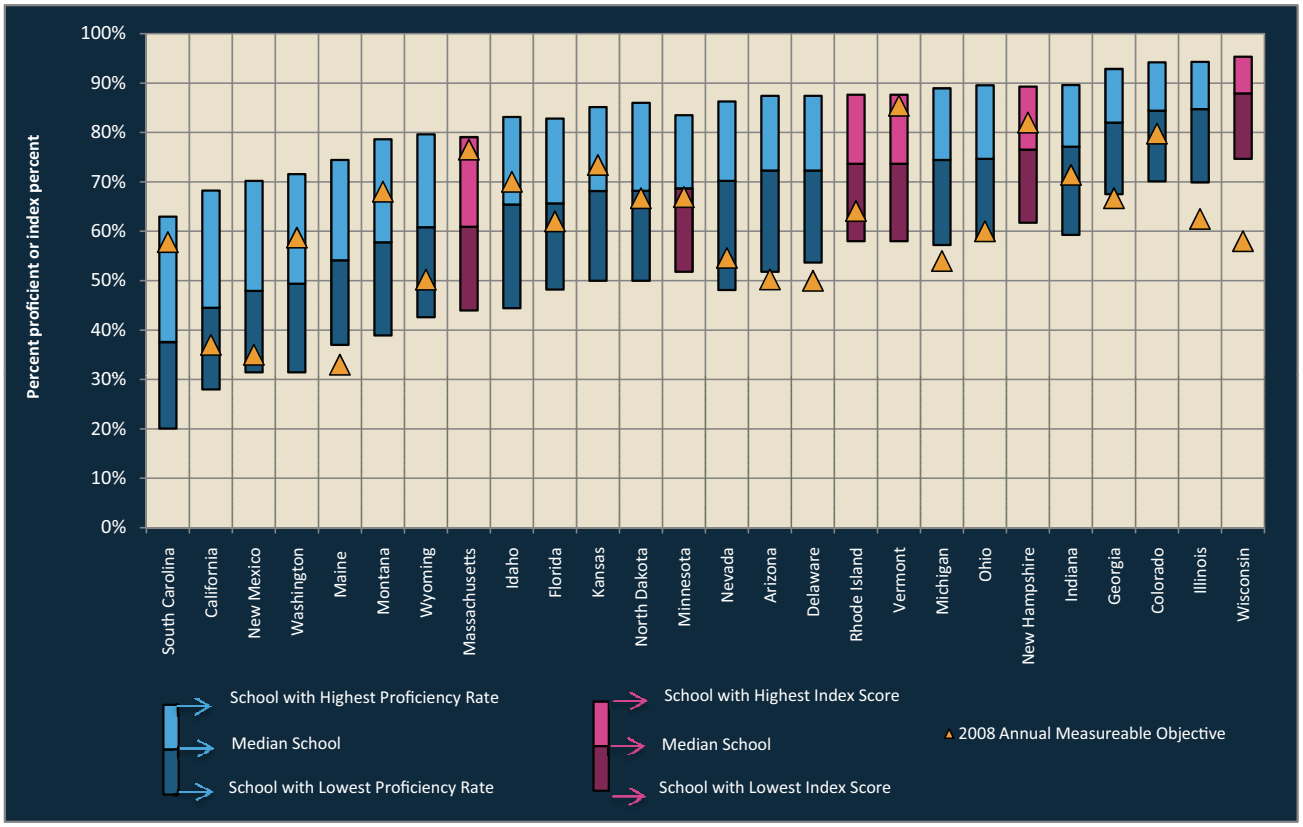

Figure 7. Math proficiency rates of the middle school sample relative to each state's 2008 AMOs

Note: Length and color of the bar represent the difference in overall performance between the highest and lowest performing school in each state. In South Carolina, for example, the lowest performing elementary school in the sample achieved an estimated proficiency rate of about 18% (as represented by the bottom of the dark blue section of the bar), the median school achieved 43% proficiency (marked by the line between the light and dark blue sections of the bar), and the highest performing school achieved 70% proficiency (shown by the top of the light blue section). States with higher cut scores are generally located at the left end and those with lower cut scores at the right. The magenta bars represent states that give students partial credit for students who achieve at lower proficiency levels. The orange triangle represents the Annual Measurable Objective, or the proportion of students required to be proficient for the 2007–2008 school year. When the triangle is below the bar, all schools in the sample met that state's AMO.

zona, Delaware, Georgia, Illinois, Maine, Michigan, and Wisconsin), all middle schools met the 2008 math AMOs (Figure 7), and in six states (Arizona, Georgia, Illinois, Michigan, Ohio, and Wisconsin), all middle schools met the reading AMOs (Figure 8). (Again, keep in mind that these results are for schools overall, not for individual subgroups.)

In a few states, however, the AMOs are very challenging. The vast majority of the sample middle schools fail to meet the math AMO in South Carolina (Figure 8). In two of the states (Massachusetts and Vermont) that use hybrid indexes, the majority also failed to meet the math AMOs (note how the AMO triangle appears at the top of each state's bar). The same is true of the reading AMOs in South Carolina, Idaho, North Dakota, Montana, and Vermont. Vermont's case is particularly interesting because it shares a common state test with Rhode Island and New Hampshire. Despite the use of a common test, more of the sample schools failed to meet the AMO in Vermont than in Rhode Island or New Hampshire because Vermont's AMO is higher.

These projections illustrate the importance of considering the AMOs in assessing the impact of NCLB. Much has been made of differences in the proficiency cut scores among the various states, but it's clear that differences in the AMOs have as much impact on the final AYP determination as the differences in cut scores. Some states with high cut scores have not set AMOs that are difficult for most schools to attain. And some states with low proficiency cut scores have AMOs that many schools would not meet. **It is the combination of these two variables that largely determines how easy or difficult it is for schools to make AYP.**

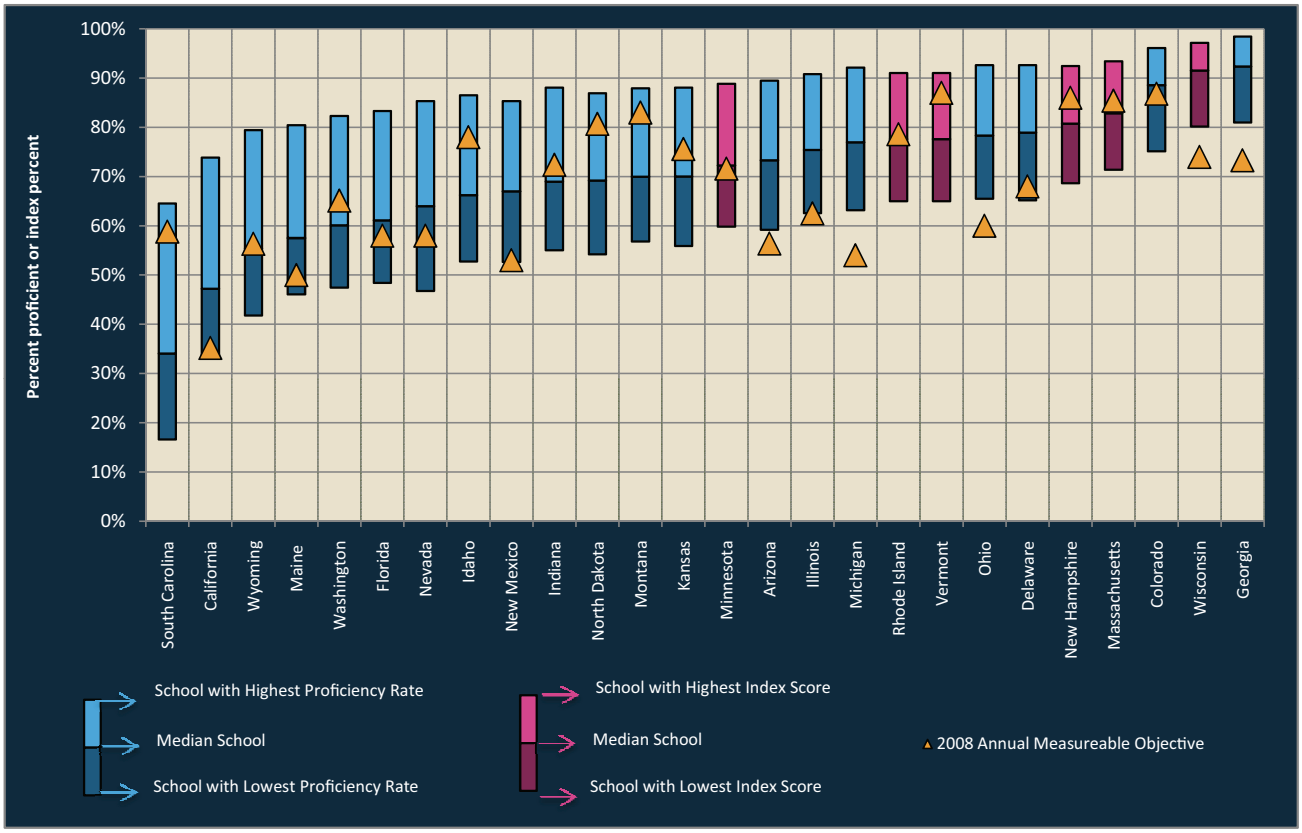

Figure 8. Reading proficiency rates of the middle school sample relative to each state's 2008 AMOs

Note: Length and color of the bar represent the difference in overall performance between the highest and lowest performing school in each state. In South Carolina, for example, the lowest performing elementary school in the sample achieved an estimated proficiency rate of about 18% (as represented by the bottom of the dark blue section of the bar), the median school achieved 43% proficiency (marked by the line between the light and dark blue sections of the bar), and the highest performing school achieved 70% proficiency (shown by the top of the light blue section). States with higher cut scores are generally located at the left end and those with lower cut scores at the right. The magenta bars represent states that give students partial credit for students who achieve at lower proficiency levels. The orange triangle represents the Annual Measurable Objective, or the proportion of students required to be proficient for the 2007–2008 school year. When the triangle is below the bar, all schools in the sample met that state's AMO.

The Lowdown on Proficiency Cut Scores and AMOs

The data for Factor 1 lead to several conclusions:

- Disparities in how high or low states set their cut scores lead to large differences in proficiency rates when these various cut scores are applied to a single sample of schools. These inconsistencies make it difficult to know what proficiency really means when comparing states to each other.

- Disparities in the AMOs further cloud interpretation of a school's AYP status. **The combination of big differences in cut scores and AMOs yields a lack of transparency across most state accountability systems.** This murkiness allows a state to correctly claim that its test is more difficult than most, while at the same time permitting nearly all schools, including poor performers, to make AYP because of low AMOs. But other states that have been criticized for their low NCLB proficiency standards (e.g., Colorado), have AMOs that seem reasonable relative to their tests. In these states, many schools may fail to meet their AMOs despite seemingly high proficiency rates.

- In a majority of cases, the math and reading AMOs for the schools' overall populations were met. Despite this, the data will ultimately show that the majority of elementary schools meeting overall proficiency targets ultimately failed to make AYP largely due to subgroup performance; the situation was similar for middle schools. We discuss this further under Factor 3.

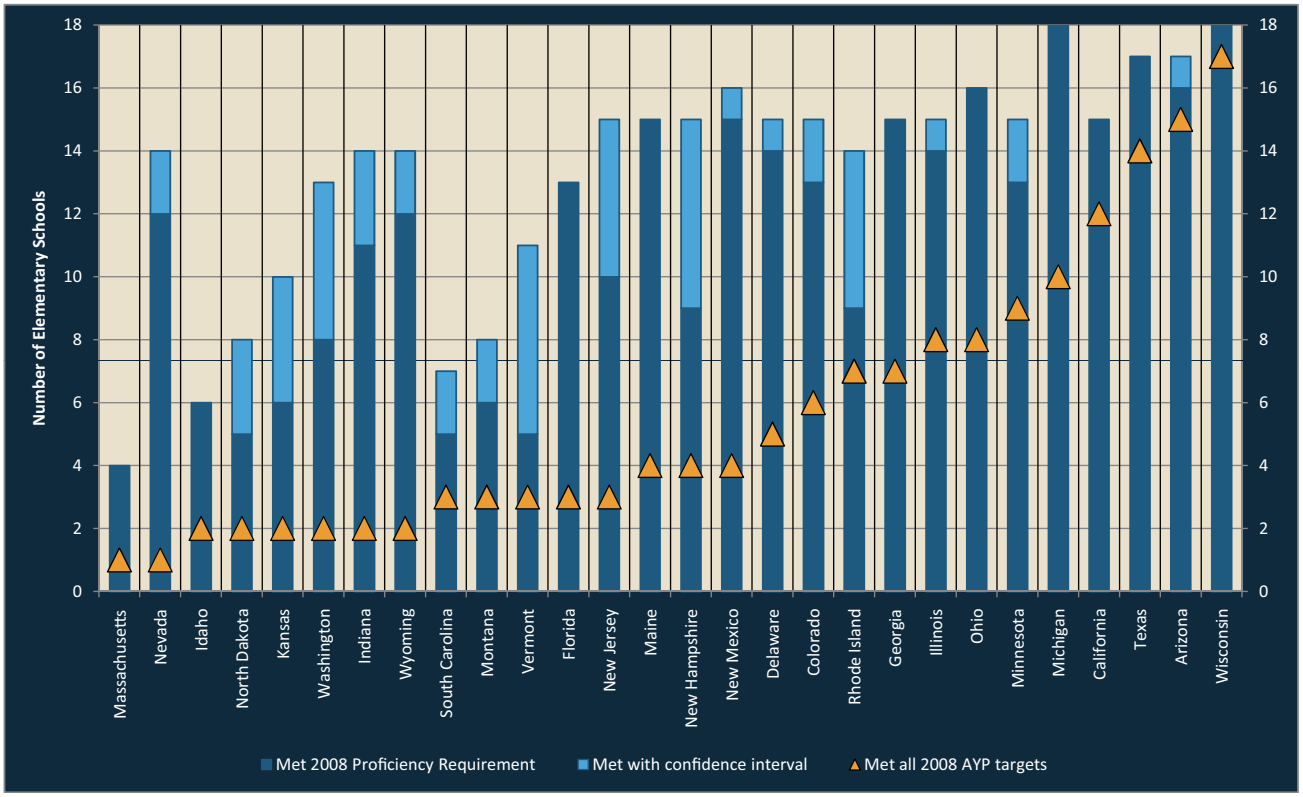

Figure 9. Number of elementary schools meeting 2008 AMOs with and without confidence intervals, by state

Note: The dark blue bars show the number of schools in each state that met their Annual Measured Objectives without employing a confidence interval. The light blue bars show the number of schools that required a confidence interval to meet the target. The orange triangles show the number of schools that ultimately made AYP (with all subgroups meeting their AMOs). For example, the figure shows that despite the fact that 14 elementary schools in Nevada met their math and reading AMOs for their overall student population– two with the help of a confidence interval–ultimately only 1 of those 14 made AYP.

How the Confidence Interval Comes into Play (Factor 2)

Nineteen of the 28 states we studied apply a confidence interval to proficiency test results. For this study, we applied the respective confidence intervals in those states that use them. Table 2 isolates the effect of the confidence intervals and shows how frequently these margins helped elementary schools meet their AMOs for their overall student populations. **In the majority of cases (63%), elementary schools met the AMO without the help of the confidence interval. The confidence interval was required to meet the AMO in about 11 % of cases, and in about 26% of the cases, schools failed to meet the AMO even with the assistance of the confidence interval.**

Table 2. Elementary school sample performance relative to AMOs with and without confidence intervals

Condition	Number of cases and percentage of total
Total measurements (18 schools X 28 states)	504
Cases meeting math and reading AMOs without confidence interval	320 (63%)
Cases meeting AMOs with confidence interval	53 (11%)
Cases not meeting AMOs (even with confidence interval)	131 (26%)

Figure 9 disaggregates the overall proficiency data to show how frequently the confidence interval helped our sample schools meet their 2008 overall proficiency targets in the various states. In 18 states at least one school benefited from the confidence interval in one or both subjects. In five states (New Hampshire, New Jersey, Rhode Island, Washington, and Vermont), five or more schools benefited from it. Overall, however, the vast majority of schools across states that met their AMOs for their overall student population did so without the assistance of a confidence interval.

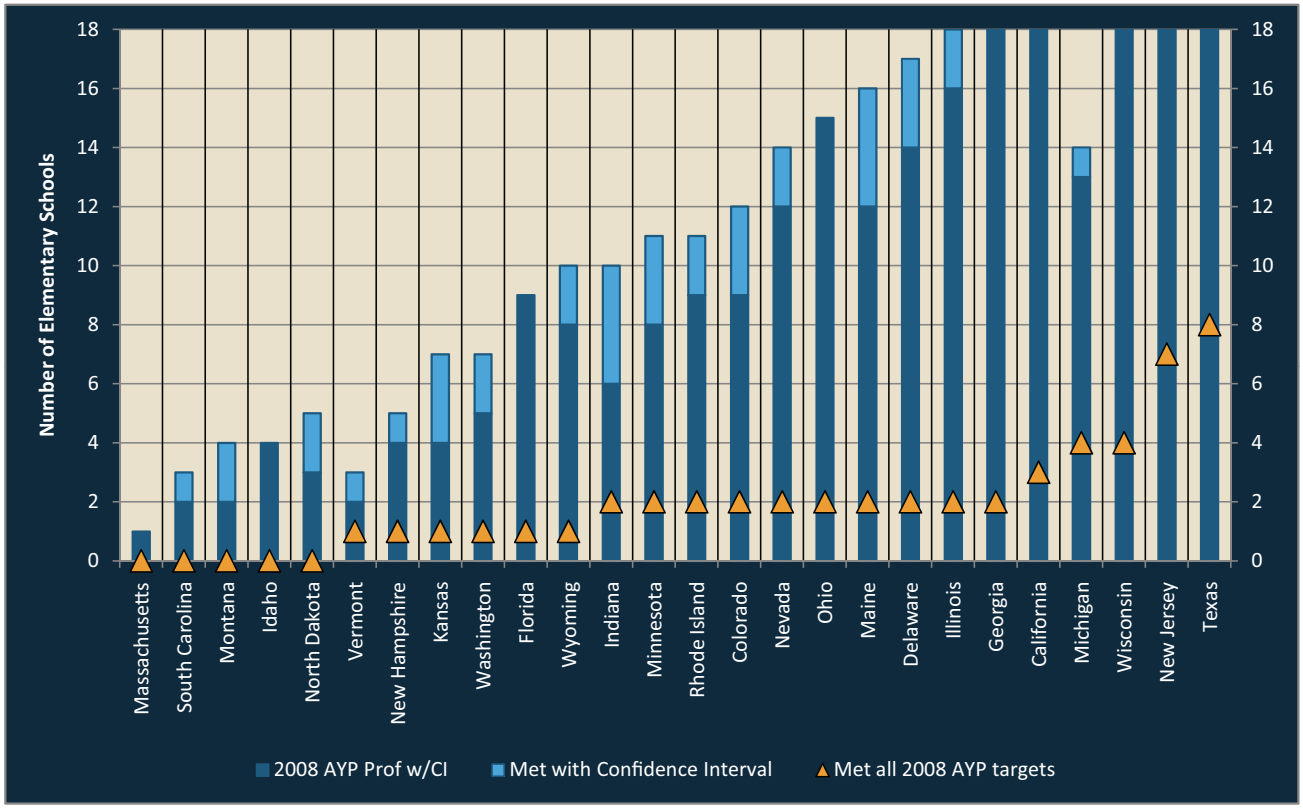

Figure 10. Number of middle schools meeting 2008 AMOs with and without confidence intervals, by state

Note: The dark blue bars show the number of schools in each state that met their Annual Measured Objectives without employing a confidence interval. The light blue bars show the number of schools that required a confidence interval to meet the target. The orange triangles show the number of schools that ultimately made AYP (with all subgroups meeting their AMOs). For example, the figure shows that despite the fact that 14 middle schools in Nevada met their math and reading AMOs for their overall student population–two with the help of a confidence interval–ultimately only 2 of those 14 made AYP.

Table 3 shows that the confidence interval was not quite as helpful to the middle school sample, since it pushed schools past their overall proficiency target in just 8% of cases. In only two states, Indiana and Maine, did the confidence interval help as many as four schools (Figure 10).

Figures 9 and 10 illustrate the effect of the confidence interval when it is applied to the overall population in our sample schools. It is important to remember, however, that when the confidence interval is used, it is not only applied to the overall student population within this study but also to all qualifying subgroups. Thus, the ultimate impact of the confidence interval is larger than the impact depicted in these two figures.

In the analyses appearing in the remainder of this report, confidence intervals were applied to all eligible subgroups in our sample schools, and the results reflect their inclusion. However, we chose not to disaggregate all figures in the report to show the confidence interval's impact because it would have added greatly to the report's length and complexity.

Table 3. Middle school sample performance relative to AMOs with and without confidence intervals

Condition	Number of cases and percentage of total
Total measurements (18 schools X 26 states*)	468
Cases meeting math and reading AMOs without confidence interval	248 (53%)
Cases meeting AMOs with confidence interval	38 (8%)
Cases not meeting AMOs (even with confidence interval)	182 (39%)

*Note: Texas and New Jersey state analyses were not conducted for the middle school sample because proficiency cut score estimates for all middle school grades were not available in these states.

The Lowdown on Confidence Intervals

To summarize our discussion of Factor 2:

- In the majority of cases, schools were able to meet AMOs for overall proficiency without the assistance of a confidence interval.
- In eight to eleven percent of cases, however, the confidence interval allowed schools to meet the AMO for their overall student population.
- When subgroups are considered, the impact of the confidence interval on ultimate AYP determinations is larger.

How the Performance of Student Subgroups Affects a School's Chances of Making AYP (Factor 3)

In this section, we discuss the impact of subgroup performance in general on AYP, including two case studies that show how the state in which a school is located impacts a school's chances of making AYP. Then we turn to a discussion of the performance of specific subgroups, namely low-income students, minority populations, LEP students, and SWDs.

Even if a school's overall proficiency rate is sufficient to meet the AMOs for math and reading, the school must also meet these same targets for each qualifying subgroup to ultimately make AYP. One consistent aspect of NCLB is that within a state, all subgroups must meet the same target. But the minimum size that qualifies a subgroup for separate evaluation differs across states. Some states require groups as small as five students to be evaluated; other states set subgroup minimums at 100 or more (see the State Reports section of this report for the particular requirements of each state).

As shown earlier, it's the combination of cut scores and AMOs that largely determines how easy or difficult it is for schools to make AYP. But a third factor, the minimum subgroup size, is also critical. **As the number of qualifying subgroups within a school increases, each new subgroup introduces another AMO that must be met.** The nature of the qualifying subgroup also makes a difference. It may be easier for a school to address poor performance in an ethnic subgroup than it is to address poor performance among SWDs, or LEP students.

The Case of Chaucer Middle School – A high performing, high growth school runs aground

Chaucer is the highest performing middle school in our sample. Table 4 summarizes the ranking of its students relative to the other middle schools in the sample. Chaucer ranks either first or second in achievement among each of the subgroups in the sample that were large enough for evaluation.

Table 4. Ranking of Chaucer middle school students relative to entire middle school sample

	Student Count	Ranking among middle school sample (reading)*	Ranking among middle school sample (math)*
All students	1118	1st	1st
Low-income students	112	1st	1st
Hispanic/Latino students	135	1st	1st
African American students	31	2nd	1st
Asian students	153	1st	2nd
LEP students	61	1st	2nd
SWDs	88	2nd	1st

* Minimum *n* of 10 students required for consideration. There are 18 middle schools in the sample.

LEP=limited English proficient; SWDs=students with disabilities

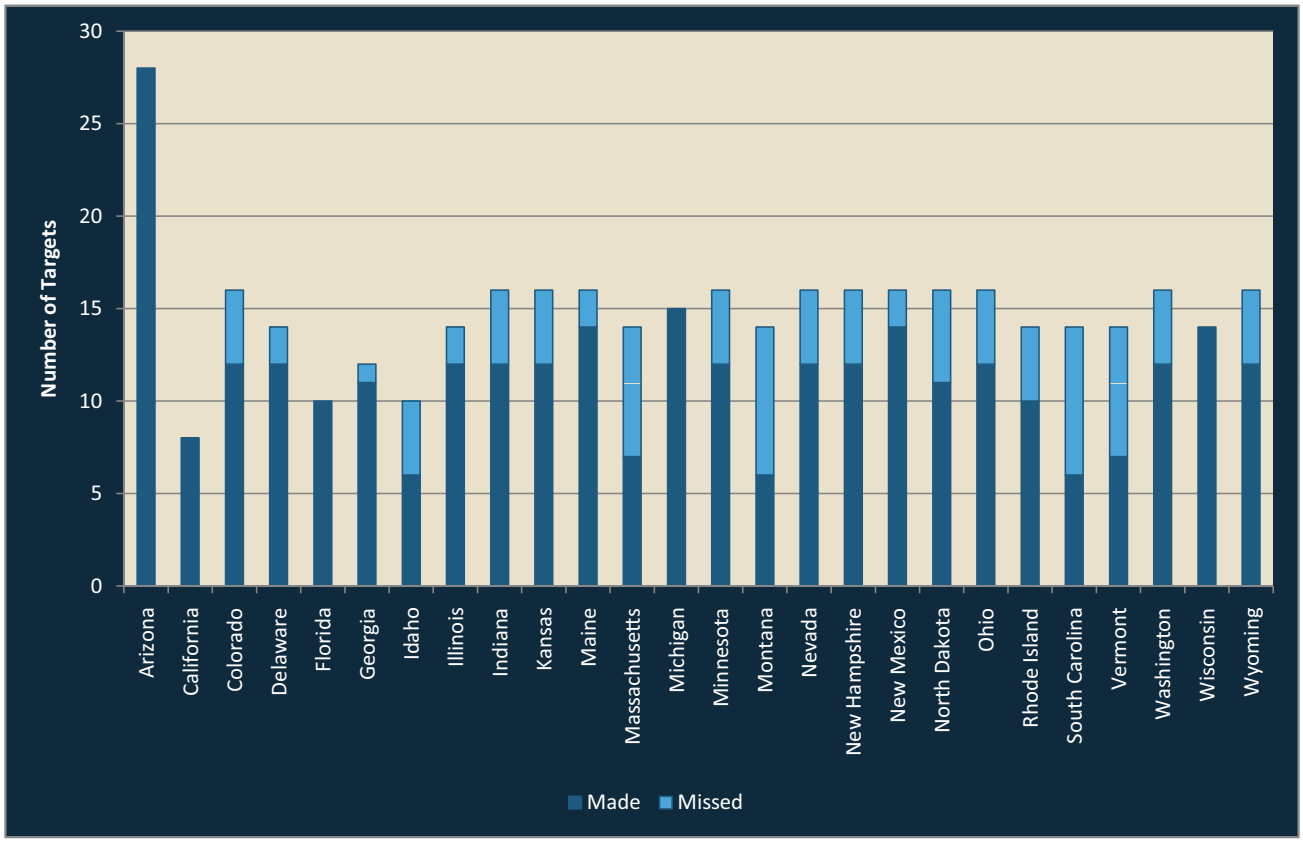

Figure 11. Number of subgroup targets met by Chaucer middle school in 2008

So how did Chaucer perform relative to the states' AYP requirements? Miserably. Chaucer made AYP in only 5 (Arizona, California, Florida, Michigan, and Wisconsin) of the 26[2] states evaluated (Figure 11). What caused this? Certainly not Chaucer's overall performance, which exceeded the annual targets in every state. Was it because of the performance of Chaucer's low-income or minority students? This is a partial explanation. Indeed, Chaucer's low-income subgroup failed to make AYP in six states and one or more of its minority subgroups failed in five states (not shown). This happened despite the fact that all of these subgroups showed above average performance relative to students in the NWEA norm group in their respective grades.

But the biggest explanation for Chaucer's failure is the performance of its LEP students and its SWDs (not shown). The LEP subgroup met its AMOs in only 2 states, failing in 20. (In the other four states, the size of this subgroup fell below the states' minimum for inclusion.) Similarly, the SWDs subgroup made its AMOs in only 2 of 26 states, failing in 21. The irony here is that Chaucer's LEP and SWD subgroups performed better than almost every other subgroup in the sample. So here is a school that is taking students with known learning challenges, presumably providing more effective help to these students than the other schools in the sample, and still failing to make AYP in more than 75% of the cases we studied. In fact, no school in the sample served students in these subgroups better. Chaucer himself aptly described the predicament of his namesake school; "…If gold rusts, what shall iron do?" If a school like *this one* is labeled a failure under NCLB, just where does one think its students should go to be better served?

In short, Chaucer ran aground primarily for two reasons. First, it's at a huge disadvantage because it's judged on

[2] While 28 states are included in the study for elementary school results, we lacked sufficient data to include Texas and New Jersey in the middle school results. Thus, middle school results are limited to 26 states.

Table 5. Ranking of Pogesto middle school students relative to entire middle school sample

	Student count	Performance rank among middle school sample* (reading)	Ranking for student growth among middle school sample* (math)
All students	54	14th	4th
Low-income students	26	3rd	5th
White students	41	18th	5th
Hispanic/Latino students	12	7th	4th

* Minimum *n* size of 10 students required for consideration. There are 18 total middle schools in the sample.

whether two subgroups with documented learning challenges—limited English proficient students and students with disabilities— met a fixed and somewhat arbitrary proficiency target, rather than whether it produced strong results and improvement in the performance of these groups. Second, it is a large school in a diverse community, which means that there are many subgroups of students and many of these groups are larger than the minimum *n* size required for evaluation. Large, diverse schools are accountable for the proficiency rate of a large number of subgroups—meaning they have many more targets to meet. On the other hand, smaller schools may be less effective, yet meet AYP because they have fewer qualifying subgroups and fewer targets to hit. Our next example illustrates this problem.

The Case of Pogesto Middle School - Small size benefits a low-performing school

Pogesto, an alternative school serving middle school students, was one of the lowest performing schools in the sample. It ranked 14th out of 18 schools in overall performance in reading and 18th in terms of white subgroup performance in reading (Table 5). Its students averaged about 3.9 scale score points below NWEA's norms, the equivalent of roughly one-half grade level. All Pogesto subgroups with counts greater than ten performed below NWEA norms. On the other hand, growth rates in math at Pogesto were above average; it performed in the top-third of the middle school sample in this regard.

Based on the results for Chaucer, we would expect Pogesto to fail to make AYP in almost every state. But Pogesto made AYP in 15 of the 26 states studied (Figure 12); only one school in the middle school sample performed better. How did this happen?

The answer is simple. With 54 students, Pogesto had fewer students than any of the other middle schools in the sample. Its subgroups are so small that one is rarely large enough to be included. In 19 of the 26 states in our study,[3] we evaluated Pogesto solely on the reading and math performance of its general student body and, in some of these states, on the performance of its white student subgroup. In only seven states (these are the states with more than four subgroup targets in Figure 12) was Pogesto required to meet AMOs with additional subgroups, and in five of these seven states, it made AYP (Arizona, Maine, Minnesota, Nevada, New Mexico).

Pogesto is not a bad school. It is actually an alternative school that serves students who have not performed well

Table 6. AYP designations for Pogesto and Chaucer middle Schools in 26 states

Both made AYP	Pogesto made AYP – Chaucer did not	Chaucer made AYP – Pogesto did not	Both failed to make AYP
4 states	11 states	1 state	10 states

[3] Recall that two states (Texas and New Jersey) were not included in the middle school analysis because of insufficient data.

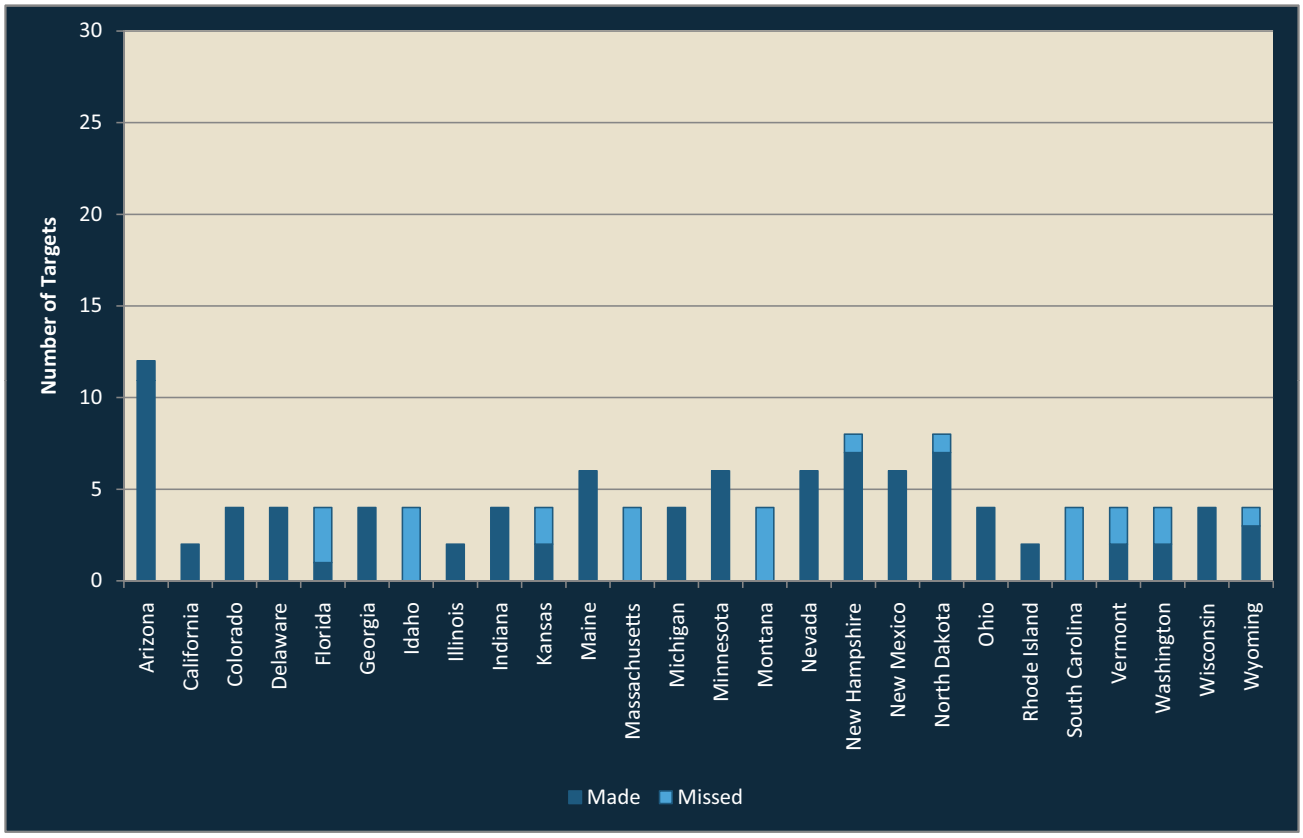

Figure 12. Number of subgroup targets met by Pogesto middle school (2008)

in other settings. Its low-income students performed near the top of the sample (though below the NWEA average) and the school's growth was within the upper third of the schools sampled. Whether Pogesto is a good or bad school, however, is not the point. Instead, the question is whether Pogesto—and other schools in the sample—are judged consistently. The answer is no. In this study, Pogesto was less effective than Chaucer by almost any measure, yet most state accountability systems have indicated otherwise. Indeed, it is remarkable that only one state (Florida) appropriately "passed" the higher performing, higher growth Chaucer while "failing" the lower performing, lower growth Pogesto (Table 6). Even more remarkable is the fact that Pogesto met AYP in 11 states where Chaucer failed to do so.

Again, Pogesto made AYP in most states because it's small and has few subgroup targets to hit, and Chaucer failed because it's large and has many subgroup targets to hit. Next, we isolate the effect of particular subgroups on the study sample.

Performance of low-income students

Even if the overall proficiency rate within a school is sufficient to meet the AMOs for math and reading, schools must still meet these same objectives for each qualifying subgroup in order to make AYP. After white students, the largest of the subgroups is typically low-income students. Table 7 summarizes the performance of this subgroup of students in the elementary school sample.

Table 7. Elementary school sample performance relative to the AMOs for low-income students

Condition	Number of cases and percentage of total
Total number of cases (18 schools X 28 states)	504
Number of cases in which low-income group was below the minimum subgroup size	55 (11%)
Number of cases in which low-income group met all AMOs	223 (44%)
Number of cases in which low-income group failed to meet one or more AMOs	226 (45%)

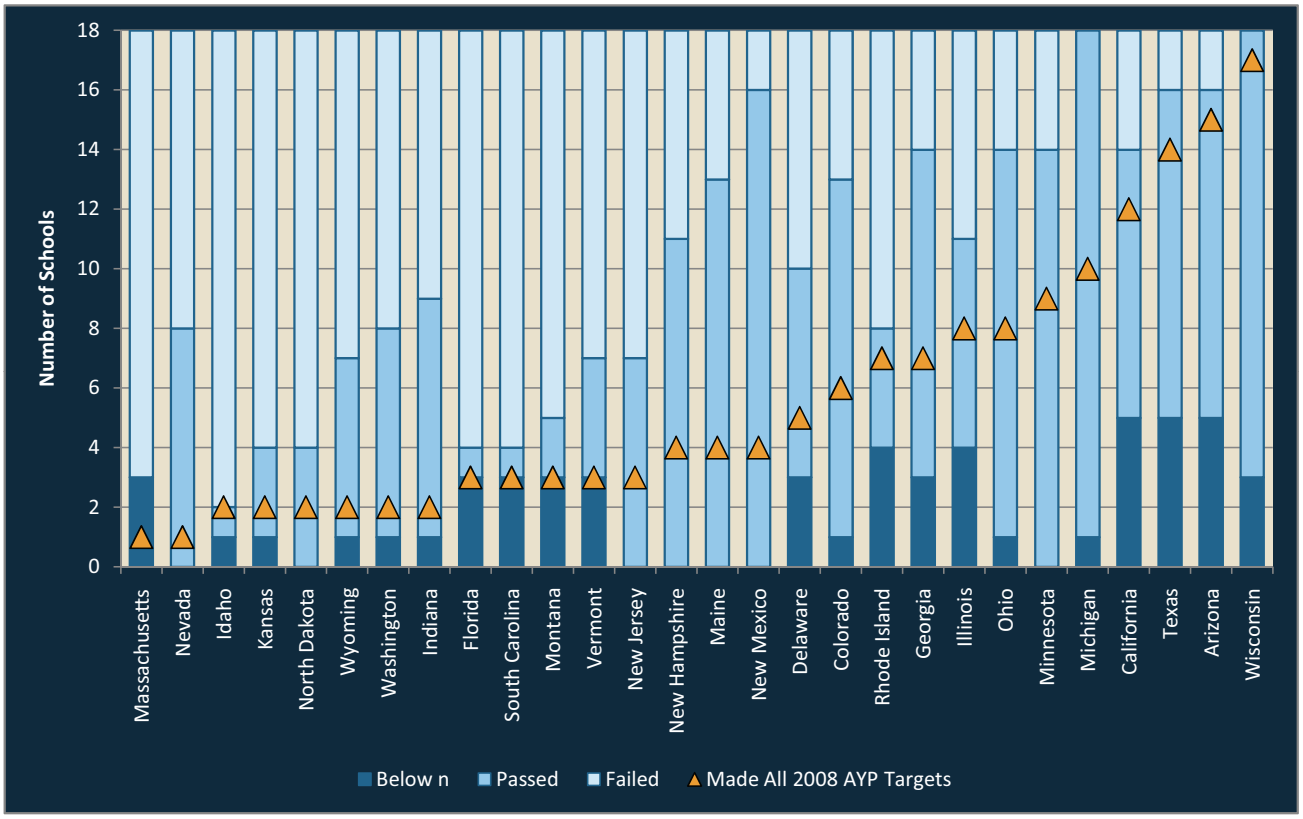

Figure 13. Number of elementary schools meeting 2008 AMOs in math and reading for their low-income student subgroup

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. For example, in Massachusetts, every elementary school with a qualifying low-income subgroup failed to meet its AMOs. In Michigan, however, every school with a qualifying low-income subgroup passed its AMO. Note, however, that even though all the low-income subgroups met their AMOs in Michigan, only 10 of the 18 schools ultimately made AYP (indicated by the orange triangle). The remaining eight failed to make AYP because of some other subgroup.

Subgroup counts were below the minimum size in only 11% of our cases. In 44% of cases, the low-income subgroup met all AMOs; it failed one or more AMO in slightly more cases (45%).

Table 8. Middle school sample performance relative to the AMOs for their low-income students

Condition	Number of cases and percentage of total
Total number of cases (26 states X 18 schools)	468
Number of cases in which low-income group was below minimum subgroup size	27 (6%)
Number of cases in which low-income group met all AMOs	149 (32%)
Number of cases in which low-income group failed to meet one or more AMOs	292 (62%)

Note: While 28 states are included in the study for elementary school results, we had insufficient data to include Texas and New Jersey in the middle school results. Thus, middle school results are limited to 26 states.

Figure 13 shows how the sample elementary schools fared by state. In one state, Massachusetts, all schools with a low-income qualifying population failed to reach their AMOs (failures are indicated by the light blue bar). In two states, Wisconsin and Michigan, we have the opposite situation; all the sample schools with a qualifying count for low-income students passed their AMOs (indicated by the median shade of blue).

Because the middle schools in our sample are considerably larger than most of the elementary schools, there were only 6% of cases in which the low-income subgroup fell below the minimum *n* size required for evaluation (Table 8). In 32% of the total cases, the school met its required AMO for the low-income subgroup, but schools failed in well over one-half (62%) of the cases.

In four states (Idaho, Massachusetts, Montana, and South Carolina), no middle school with a qualifying

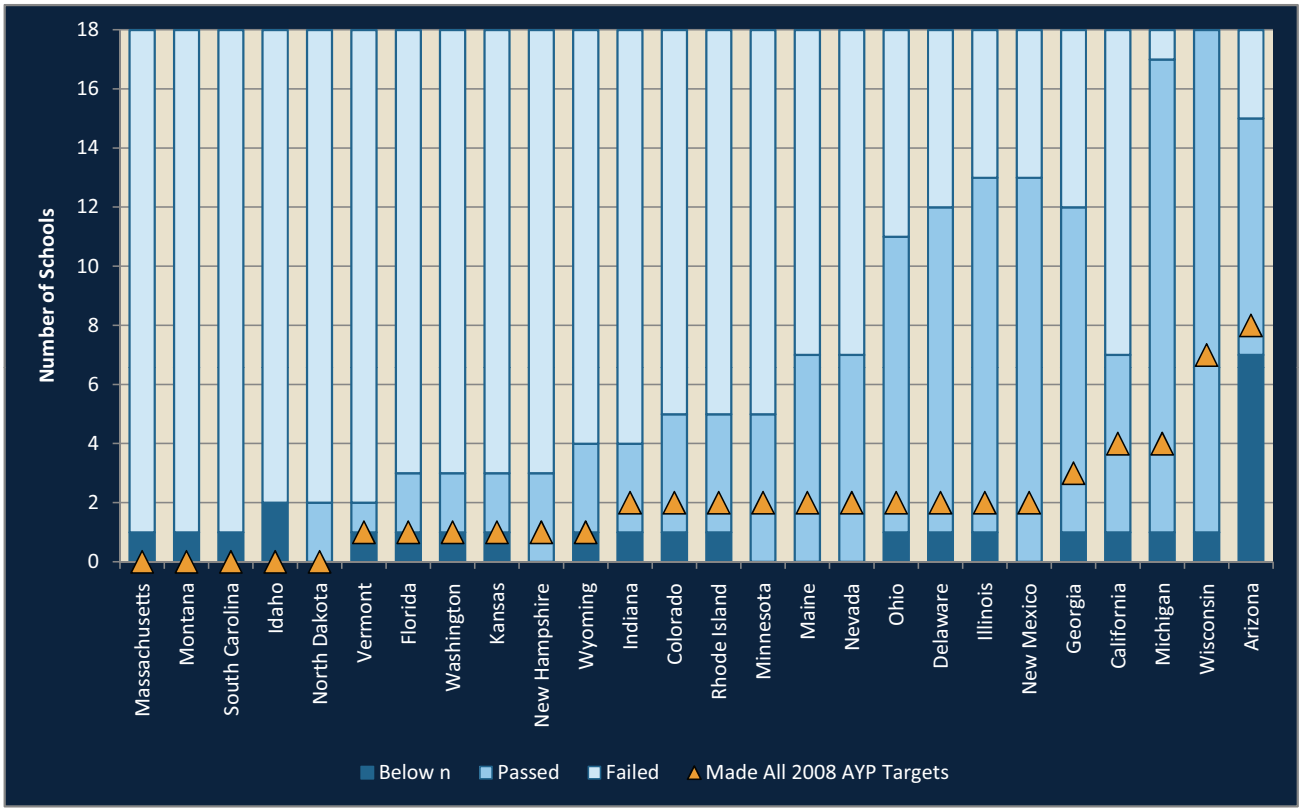

Figure 14. Number of middle schools meeting 2008 AMOs in math and reading for their low-income student subgroup

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. For example, in Massachusetts, every middle school with a qualifying low-income subgroup failed to meet its AMOs. In Wisconsin, however, every school with a qualifying low-income subgroup passed its AMO. Note, however, that even though all the low-income subgroups met their AMOs in Wisconsin, only 7 of the 18 schools ultimately made AYP (indicated by the orange triangle). The remaining 11 failed to make AYP because of some other subgroup.

low-income population met the AMOs for that group (Figure 14). There was one state, Wisconsin, in which all sample middle schools with a low-income qualifying population passed. In 18 states, half or more of the low-income subgroups within the middle school sample failed this AMO (note all of the long light blue bars in Figure 14). The AYP performance of the schools provides an interesting contrast. They show, for example, that even in states where the low-income students made their AMO, it did not necessarily help assure a positive final outcome for the school. For example, 13 schools in New Mexico met the AMO for low-income students, and 11 of the 13 still failed to make AYP.

Overall, elementary schools failed to meet the annual targets for the low-income subgroup in 45% of cases, while middle schools failed to meet it in 62% of cases. These failures were not evenly spread across states, but concentrated among about two-thirds of the sample states.

Performance of minority students

Table 9 reports the performance of minority students within the sample elementary schools relative to their 2008 AMOs for reading and math across all states studied. In about 27% of the total cases, schools in the sample had no minority group large enough to meet the

Table 9. Elementary school sample performance relative to the AMOs for their minority students

Condition	Number of cases and percentage of total
Total number of cases (18 schools X 28 states)	504
Number of cases in which all minority groups were below minimum subgroup size	134 (27%)
Number of cases in which all minority groups met all AMOs	139 (28%)
Number of cases in which some minority groups failed to meet one or more AMOs	231 (46%)

Note: Percentages may not add to 100 due to rounding.

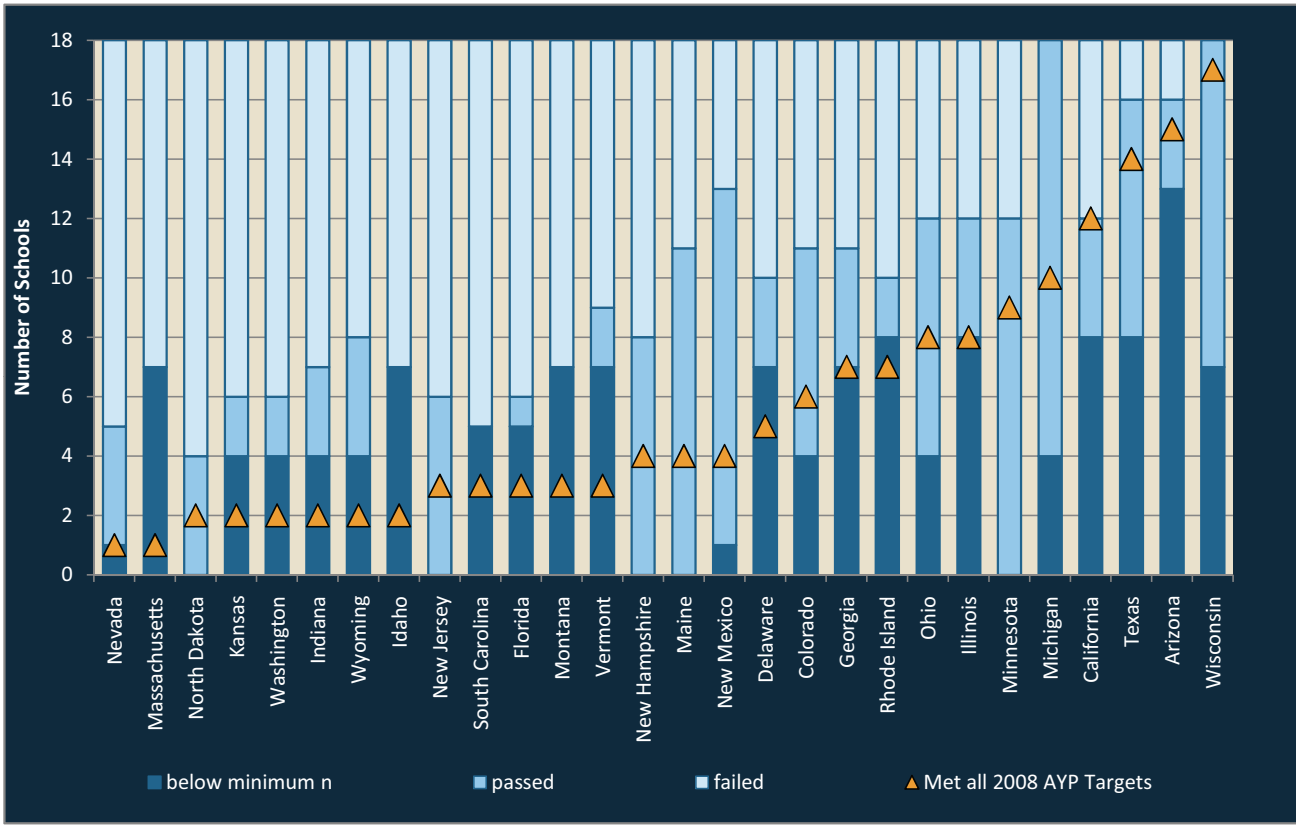

Figure 15. Number of elementary schools in which minority students met their 2008 AMOs

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. For example, in Massachusetts, every school with a qualifying minority subgroup failed to meet its AMO. In Michigan, however, every school with a qualifying minority subgroup passed its AMO. Note, however, that even though all the minority subgroups met their AMOs in Michigan, only 10 of the 18 schools ultimately made AYP (indicated by the orange triangle). The remaining 8 failed to make AYP because of some other subgroup.

minimum reporting requirement. Among the remainder, all qualifying minority groups met their objectives in math and reading in 28% of cases, but in 46% of cases, one or more minority groups failed to meet the objectives in one or both subjects.

Figure 15 shows the distribution of results for the elementary school sample by state. Because of a low minimum *n* size requirement, there were five states in the sample (Maine, Minnesota, New Hampshire, New Jersey, and North Dakota) in which all schools had at least one minority subgroup that exceeded the minimum subgroup size.

There were four states (Idaho, Massachusetts, Montana, and South Carolina) in which all schools with a minority subgroup that met the minimum *n* size failed one or more AMOs. All four of these states had relatively high cut scores. In 13 other states, more than half the schools had at least one minority group that failed to meet an annual target; these states also had cut scores that fell in the upper half in difficulty. But there were also two states, Michigan and Wisconsin, in which all schools

Table 10. Middle school sample performance relative to the AMOs for minority students

Condition	Number of cases and percentage of total
Total number of cases (26 states X 18 schools)	468
Number of cases in which all minority groups were below minimum subgroup size	40 (9%)
Number of cases in which all minority groups met AMO	103 (22%)
Number of cases in which some minority groups failed to meet one or more AMOs	325 (69%)

Note: While twenty-eight states are included in the study for elementary school results, we had insufficient data to include Texas and New Jersey in the middle school results. Thus, middle school results are limited to twenty-six states.

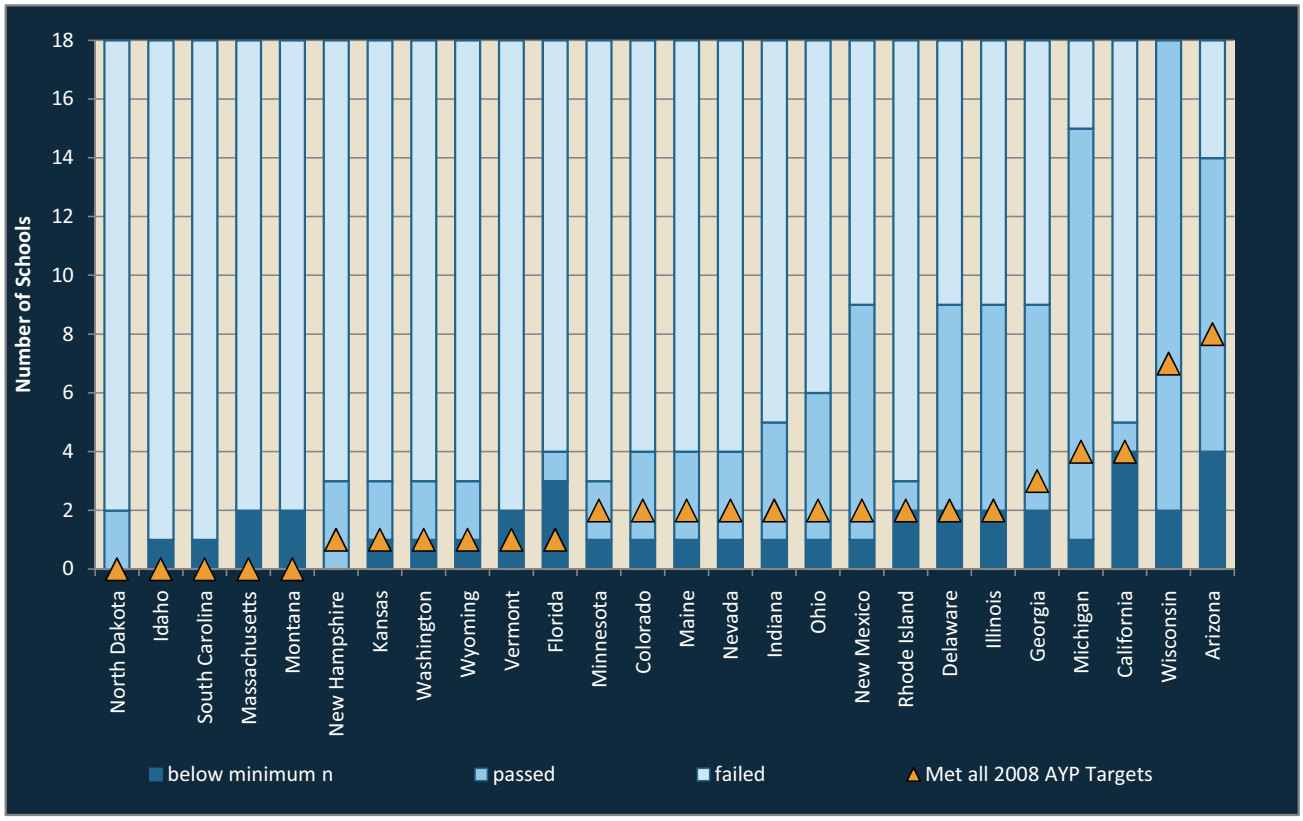

Figure 16. Number of middle schools in which minority students met their 2008 AMOs

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. For example, in North Dakota every school with a qualifying minority subgroup failed to meet its AMO. In Wisconsin however, every school with a qualifying minority subgroup passed its AMO. Note, however, that even though all the minority subgroups met their AMOs in Wisconsin, only 7 of the 18 schools ultimately made AYP (indicated by the orange triangle). The remaining 11 failed to make AYP because of some other subgroup.

with a qualifying minority group passed. These two states have both lower than average cut scores and lower than average AMOs. Finally, there are several states in which many schools that met the AMOs for their minority students ultimately failed to make AYP on some other basis. In Maine, for example, there were 11 schools in which all minority subgroups met the AMO, yet only 4 of these schools ultimately made AYP. While all schools in Michigan with a qualifying minority subgroup saw those subgroups meet the AMO, 8 of the schools failed to make AYP because of some other subgroup.

Once again, the middle schools in the sample performed worse than the elementary schools. Because middle schools are generally larger than elementary schools, in just 9% of the cases were there no minority groups in a school large enough to qualify as a subgroup—less than half what was found in the elementary school group. Minority groups passed all of their proficiency objectives in 22% of cases, but failed in 69% of cases, a failure rate 22 percentage points higher than the elementary school failure rate (Table 10).

In five of the states (Idaho, Massachusetts, Montana, South Carolina, and Vermont), all middle schools with a qualifying minority group failed to meet that group's targets (Figure 16). In 19 of the 26 states, more than half the middle schools in the sample failed to meet their targets for one or more of their minority groups. The only state in which all schools with a qualifying minority group passed was Wisconsin, but more than half of the schools also passed the targets in Michigan and Arizona. Once again, there are several states in which the minority subgroups of many schools met their AMO, yet the vast majority of schools still ultimately failed to make AYP. In Michigan, for example, all minority subgroups passed in fifteen schools, but only four of these schools ultimately made AYP (indicated by the orange triangle). In Wis-

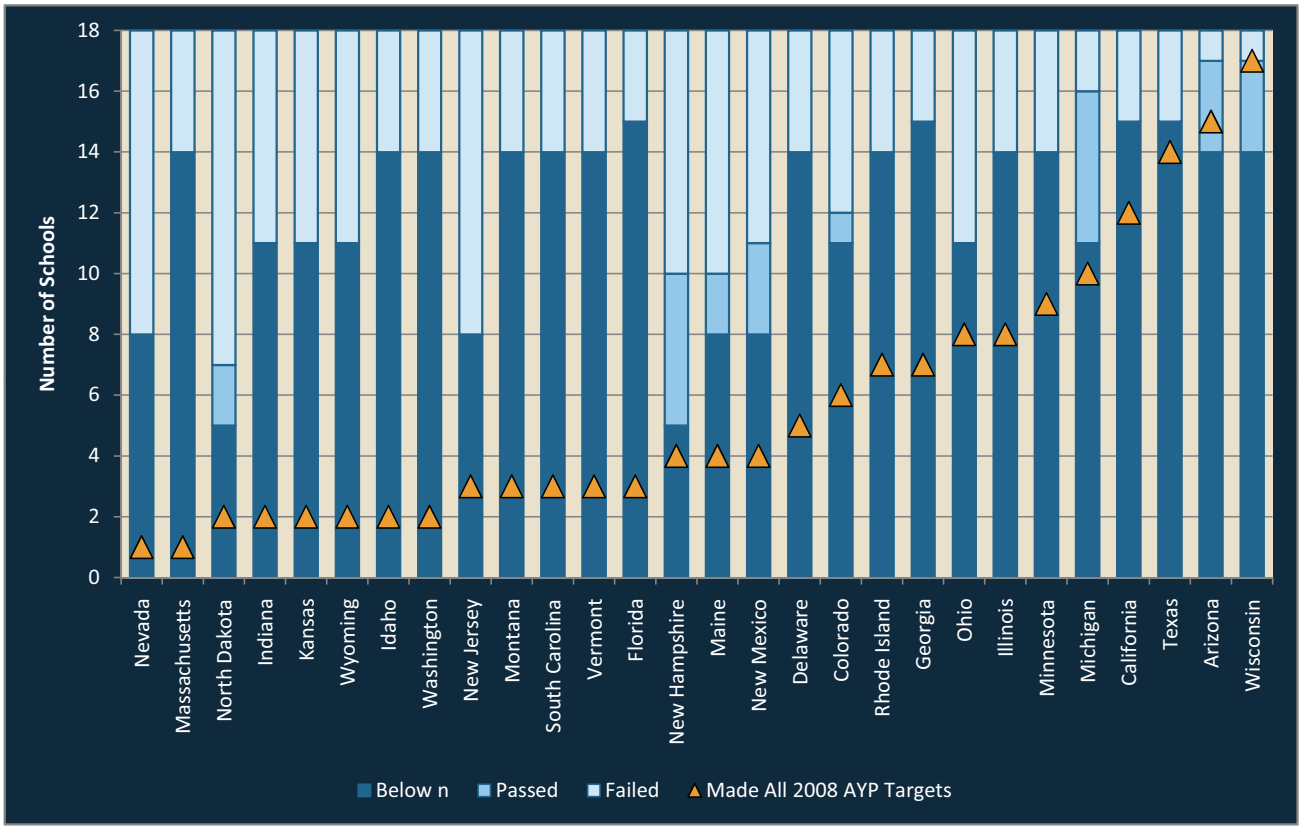

Figure 17. Number of elementary schools in which LEP students met their 2008 AMOs

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. For example, in Ohio every elementary school with a qualifying LEP subgroup failed to meet its AMO. In New Hampshire, however, five schools did not meet subgroup requirements and five schools met LEP targets (dark blue and median blue bars). However, even though ten schools met their LEP targets in New Hampshire, only 4 of the 10 schools ultimately made AYP (indicated by the orange triangle). The remaining 6 failed to make AYP because of some other subgroup

consin, all minority subgroups passed in sixteen schools, yet only seven ultimately made AYP.

Performance of LEP students

In general, LEP students are required to participate in state testing for purposes of determining AYP. Students who are not English proficient and are new to the United States need not participate in state testing during the first calendar year in which they're enrolled. Until recently, students who graduated from LEP status by achieving English proficiency were moved out of the subgroup during the year that they became proficient. In practice, this created a churning effect, in which successful students were removed from the LEP subgroup and new English language learners moved in. A mid-course change to NCLB regulations by the U.S. Department of Education now allows states to retain in the LEP subgroup, for up to two years, students who have become proficient in English. This reduces, but does not eliminate, the churning effect.

Many of the elementary schools in the sample (67% of cases) did not have LEP populations large enough to meet

Table 11. Elementary school sample performance relative to their 2008 AMOs for students with limited English proficiency

Condition	Number of cases and percentage of total
Total number of cases (18 schools X 28 states)	504
Number of cases in which the LEP group was below the minimum subgroup size	336 (67%)
Number of cases in which the LEP group met all AMOs	24 (5%)
Number of cases in which the LEP group failed to meet one or more AMOs	144 (27%)

Note: Percentages may not add to 100 due to rounding.

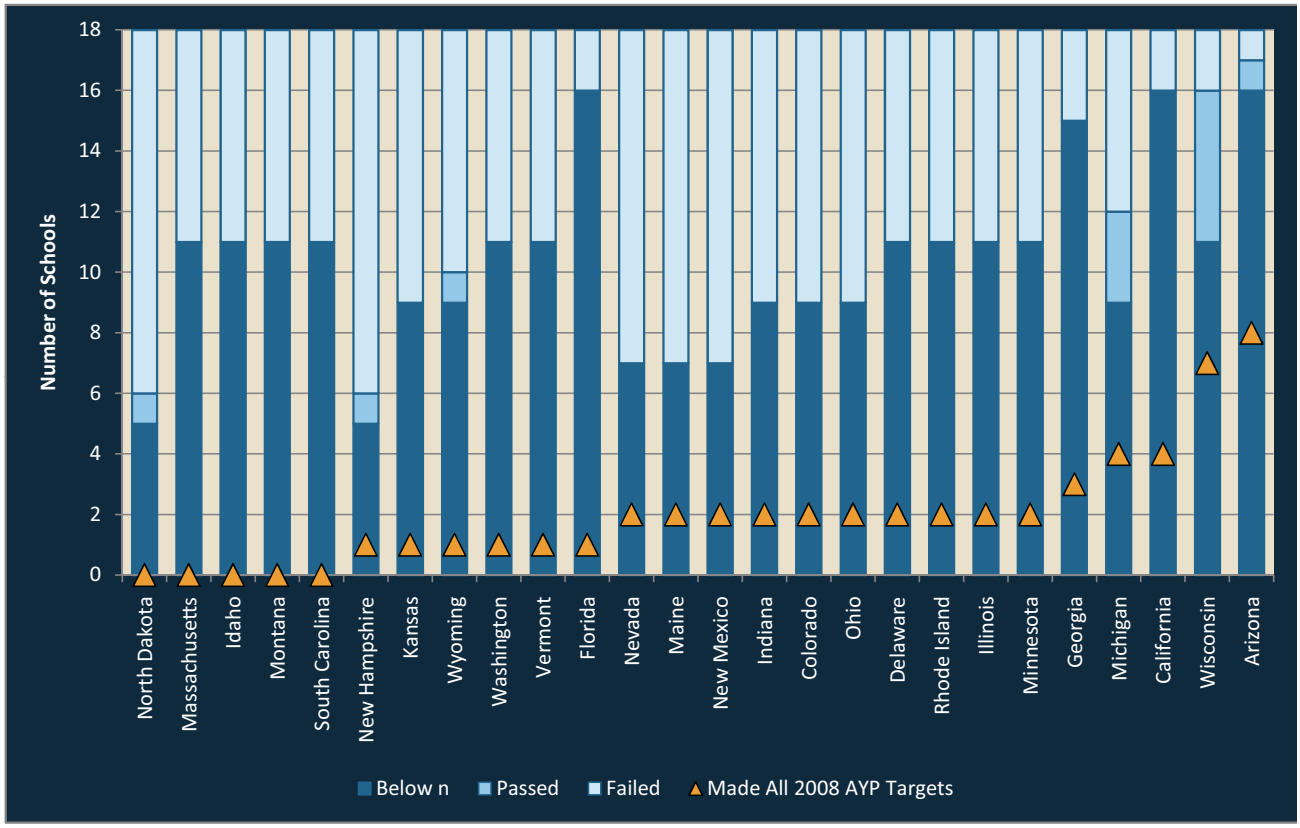

Figure 18. Number of sampled middle schools in which LEP students met their 2008 AMOs

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. In the vast majority of states (New Mexico, Indiana, Colorado, Delaware, etc.), every school with a qualifying LEP subgroup failed to meet its AMO.

the minimum *n* size in the states studied (Table 11). In situations where this subgroup's performance is counted, however, nearly all schools failed to meet their AMOs. Schools failed in 27% of total cases, nearly six times the number of cases in which schools succeeded (5%). In 20 of the states studied, all schools whose LEP population exceeded the minimum *n* size failed to meet their AMOs (indicated by the absence of a median blue bar in Figure 17).

The middle schools, again, did not perform as well as the elementary schools. Although the majority (57%) did not have LEP subgroups large enough to qualify for evaluation, a school with a qualifying count passed its AMOs in only 3% of the total cases and failed in 40% of the total cases (Table 12). In 20 of the 26 states, all schools with qualifying LEP populations failed to meet their AMOs for this subgroup (Figure 18).

Sadly, the best way to for a school to avoid failure with its LEP students is to avoid having many of them. In fact, more than half of the sample was not evaluated on the performance of these students because they fell below the various states' minimum *n* size requirements (Table 12). And nearly all of those schools that did have a qualifying LEP subgroup failed to meet the AMOs for this group.

Table 12. Middle school sample performance relative to their 2008 AMOs for LEP students

Condition	Number of cases and percentage of total
Total number of cases (26 states X 18 schools)	468
Number of cases in which the LEP group was below the minimum subgroup size	269 (57%)
Number of cases in which the LEP group met all AMOs	12 (3%)
Number of cases in which the LEP group failed to meet one or more AMOs	187 (40%)

Note: While twenty-eight states are included in the study for elementary school results, we had insufficient data to include Texas and New Jersey in the middle school results. Thus, middle school results are limited to twenty-six states.

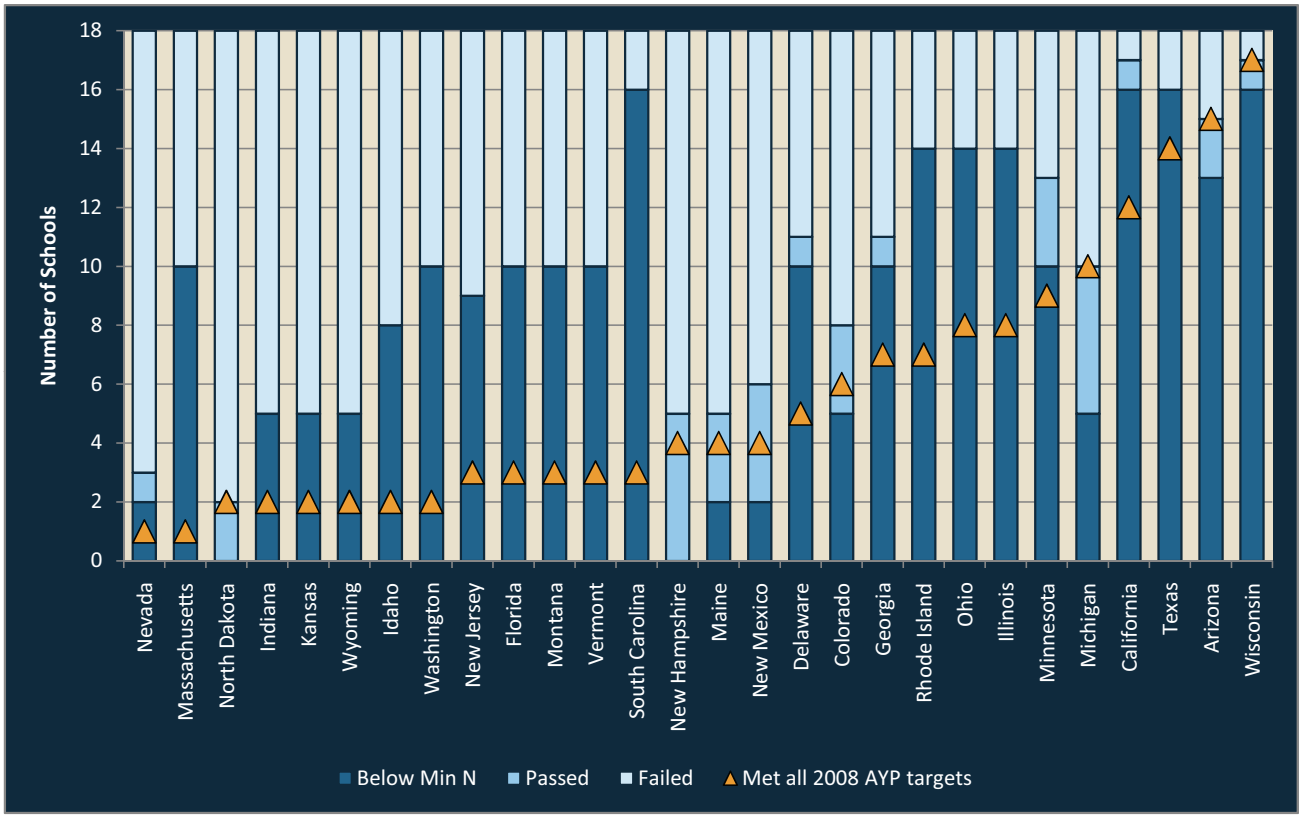

Figure 19. Number of sampled elementary schools in which SWDs met their 2008 AMOs

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. In the vast majority of states (Wyoming, Idaho, Washington, Vermont, etc.), every school with a qualifying SWD subgroup failed to meet its AMO.

Performance of SWDs

This was the final factor considered. Students with disabilities are not exempt from the NCLB 100% proficiency requirement, but states are allowed to exclude from testing up to one percent of students who have significant cognitive disabilities. States are also allowed, under a change to the NCLB regulations, to test another two percent of students using an alternative assessment.[4]

Table 13. Elementary school sample performance relative to their 2008 AMOs for students with disabilities

Condition	Number of cases and percentage of total
Total number of cases (18 schools X 28 states)	504
Number of cases in which the SWD group was below the minimum subgroup size	247 (49%)
Number of cases in which the SWD group met AMOs	32 (6%)
Number of cases in which the SWD group failed to meet one or more AMOs	225 (45%)

How does the SWD subgroup perform? Within the elementary school sample, the count of disabled students fell below the minimum *n* size in just under half of all cases (49%) (Table 13). There were 225 cases of subgroups failing to meet AMOs (45%) and only 32 cases (6%) in which the subgroups met their AMO. In fifteen states, all elementary schools whose SWD subgroup met the required minimum *n* size failed to meet their AMOs (Figure 19).

[4] Participating schools in this study did not report to us whether each student's achievement level was attained on the state's general assessment or on the alternative assessment, so we caution that some students included in these results could be eligible to take a state's alternate assessment or excluded from testing entirely. However, it's not general practice for schools to test students with severe cognitive disabilities on the NWEA assessment, so it is unlikely that these students are included here.

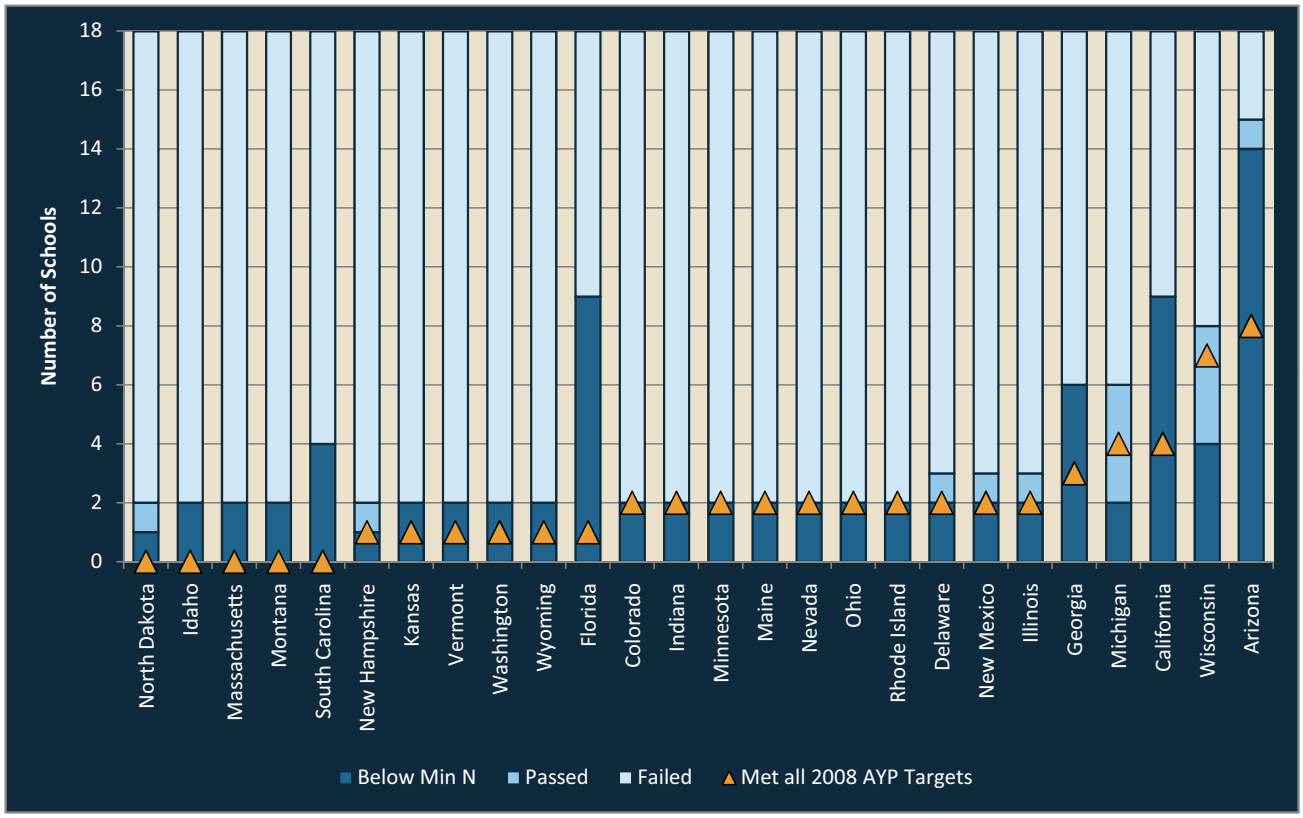

Figure 20. Number of sample middle schools in which SWDs met their 2008 AMOs

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. In the vast majority of states (Wyoming, Idaho, Rhode Island, Vermont, etc.), every school with a qualifying SWD subgroup failed to meet its AMO.

Among the middle school sample, in only 18% of cases did schools not have SWD subgroups large enough to qualify for evaluation (Table 14). Of the remaining cases where schools did have large enough SWD subgroups, middle schools met their AMOs in 3% of cases and failed to meet their AMOs in 79% of cases. In 18 of the states, no middle school surpassing the minimum *n* size met its AMO target for SWDs (Figure 20).

Table 14. Performance of the sampled middle schools relative to the 2008 AMOs for SWDs

Condition	Number of cases and percentage of total
Total number of cases (26 states X 18 schools)	468
Number of cases in which the SWD group was below the minimum subgroup size	84 (18%)
Number of cases in which the SWD group passed AMO	14 (3%)
Number of cases in which the SWD group failed one or more AMOs	370 (79%)

Note: While twenty-eight states are included in the study for elementary school results, we had insufficient data to include Texas and New Jersey in the middle school results. Thus, middle school results are limited to twenty-six states.

As with LEP students, nearly all of the schools in the sample that have SWD subgroups exceeding the minimum count failed. Because middle schools are generally larger than elementary schools, there are far more cases in which the middle school sample is evaluated (82%) than in the elementary schools (51%).

The Lowdown on Subgroup Performance

Figure 21 provides a very interesting summary of how subgroup performance affects the prospects for making AYP within our sample. Essentially it shows that schools had much more success with their low-income and minority subgroups than with their LEP and SWD subgroups. The graphic also shows that elementary schools

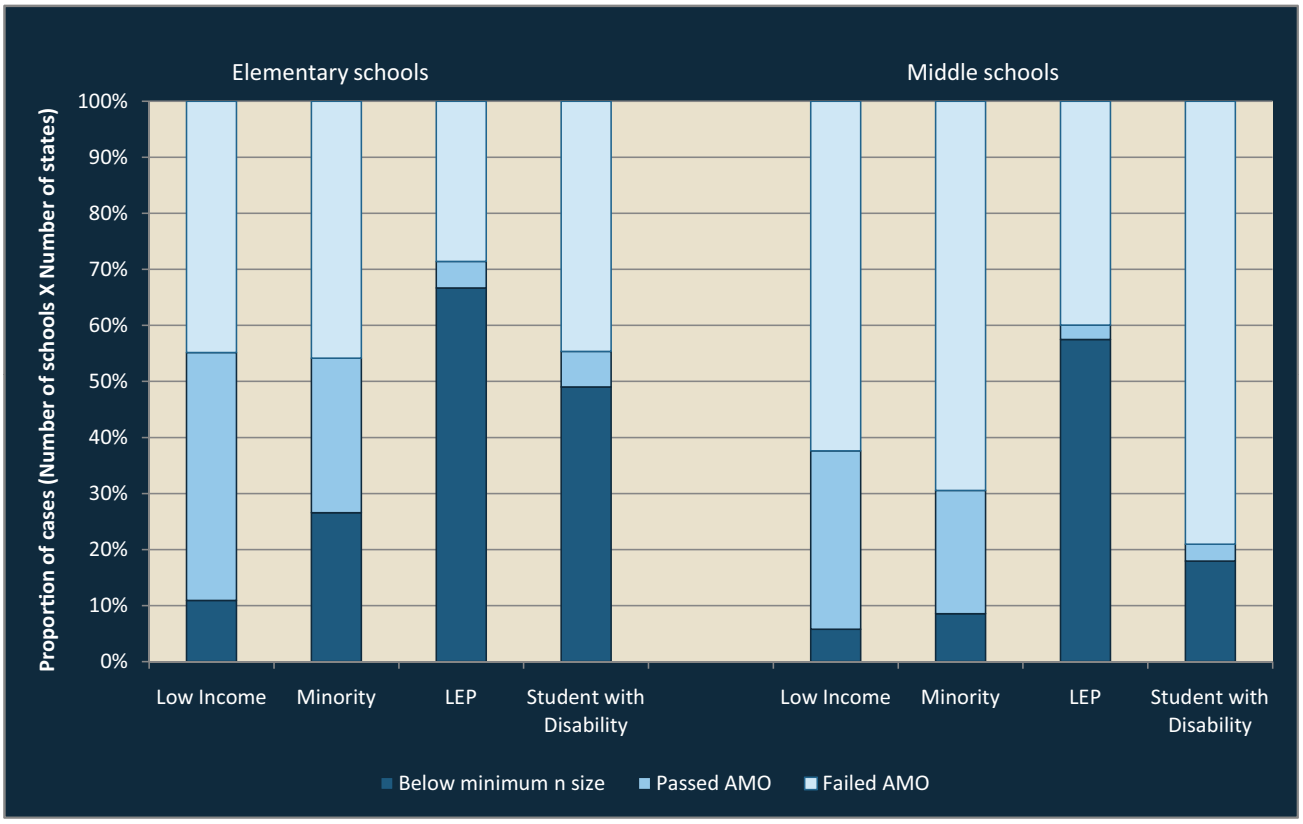

Figure 21. Summary of subgroup performance relative to AMOs

Note: The dark blue bars show schools whose count was below the minimum *n* size requirement; the median blue bars show the schools making AYP; the light blue bars show schools failing to make AYP. The figure shows that schools had much more success with their low-income and minority subgroups than with their LEP and SWD subgroups. It also shows that elementary schools failed to meet their AMOs with far less frequency than middle schools, primarily because elementary schools had far fewer subgroups that met the minimum subgroup size.

Abbreviations: SWDs = students with disabilities; AMO = annual measurable objective (yearly target)

failed their AMOs with far less frequency than middle schools, primarily because elementary schools had far fewer subgroups that met the minimum subgroup size.

While the low passing rates of low-income and minority subgroups may be frustrating, the passing rates for schools with qualifying LEP or SWD subgroups are simply astounding (as shown by the sliver of median blue in these categories in Figure 21). In the vast majority of cases, a school with a qualifying subgroup in one of these two categories failed to meet the relevant AMOs and thus failed to make AYP.[5] **The difficulty of the states' cut scores and AMOs were largely irrelevant in these cases. These subgroups failed whether the cut scores were high or low and whether the AMOs were strict or generous.**

So, to summarize:

- A state's minimum subgroup size (or *n* size) determines the number of subgroups that must meet an AMO. Since failing a single AMO causes a school to fail to make AYP, having more subgroups increases the number of opportunities for failure. This is the case with middle schools in the sample—they don't fare worse because they are less effective in educating students, but because they have more subgroups.

[5] We should note that this study may underestimate the performance of students in the LEP and SWD subgroups, mostly because of the likely differences between how LEP students and SWDs are treated in MAP, the assessment we used in this study, and in the various state standardized tests. Specifically, the U.S. Department of Education has issued new NCLB guidelines in recent years that exclude small percentages of LEP students and SWDs from taking the state test or that allow them to take alternative assessments. In this study, however, no valid MAP scores were omitted from consideration.

- Rather than claim that large schools face a "diversity penalty," it may be fairer to say that small schools enjoy a "homogeneity bonus." Small schools typically do not have to meet objectives for many subgroups since they don't have enough low income, minority, LEP or SWD students to qualify for evaluation. In large schools, these subgroups often fail to meet their AMOs (as shown in Figure 21). Because there's no reason to believe that pupils in small-school subgroups are performing at levels way beyond those in larger-school subgroups, small schools are probably fortunate that they're not accountable for these groups separately. They clearly have an easier time making AYP than larger schools.

- As indicated above, middle schools in the sample fared more poorly than elementary schools. In only 32% of cases did low-income student subgroups in middle schools meet their AMOs. Contrast this with elementary schools, where 44% of low-income subgroups met their AMOs. The picture is much the same for minority subgroups. In 22% of middle school cases, all minority student subgroups met their AMOs; the same is true in 28% of elementary school cases.

- Even more damaging to a school's chances of making AYP is the presence of a qualifying subgroup of LEP students or SWDs. In only 3% of middle school cases and 5% of elementary school cases did a LEP subgroup meet its AMOs. Similarly, in only 3% of middle school cases and 6% of elementary school cases did a subgroup of SWDS meet its AMOs. As a result, most schools that actually made AYP by our estimate did so because their LEP and SWD subgroups were too small to qualify for evaluation.

Limitations

The purpose of this study was to explore how key elements of NCLB, in this case proficiency cut scores, proficiency rate targets (AMOs), subgroup sizes, and confidence intervals may interact to affect the AYP status of schools. We hoped to shed light on such questions as "Would a school with a population and performance mix that makes AYP in California also be likely to make AYP in New Hampshire, Washington, or South Carolina?"

A sample of real schools was chosen for the study in an effort to assure a meaningful connection between our analysis and the actual conditions faced by schools. (Each school is identified by a pseudonym.) We hope this makes the study useful, informative, and interesting. This study literally shows what happens when you take the performance of a set of schools on a single assessment, estimate different proficiency cut scores for that assessment based on a sound estimate of the difficulty of the standards in different states, and apply the AYP rules in place for that state to the dataset. This kind of illustration is very useful when one wants to evaluate whether the effect of the NCLB accountability policy is likely to be consistent across states. And that was our purpose here.

We must emphasize, however, that the MAP assessment and analytic tools will not precisely replicate the sample schools' performance on their state tests. While all students in the sample took some form of their state assessment, schools did not identify whether students took the regular assessment or the alternative assessment. For the purposes of our study, a student's performance on the various states' assessments was projected from their MAP scores. Therefore, it is possible that some students we identify as failing, particularly LEP students or students with disabilities, would be eligible to take the alternative form of the assessment

in some states. We have no data that allow us to predict how these students might have performed on the alternative assessment.

Some students within a school who participated in state testing did not participate in MAP testing (and vice versa), but we included only students who participated in both MAP and state tests in our sample. As a result, the students included for estimation in our study were not identical to the students who participated in state testing that same school year. Tables A-4 and A-5 (in Appendix A) show differences in the count of students taking MAP and their state test and those who participated only in their state test for the sample schools. For all but two of the sample schools, the MAP results predicted, within five percentage points, the school's actual performance on their state test. In addition, our pilot study (Cronin et al. 2007b) found that the rates of proficiency estimated on the MAP assessment for samples of students closely paralleled the rates of proficiency reported on state tests.

In testing the effects of confidence intervals, we followed the methodology employed by the state in their calculations. Because MAP is an adaptive assessment[6] (state tests are generally fixed form), our estimate of the confidence intervals associated with MAP may be narrower in some states than the confidence interval associated with the state assessment. This happens because the standard error of measure associated with MAP is generally smaller for very high and low performing students than the standard error of measure on a fixed form test. In these circumstances, our confidence interval calculation may slightly understate the actual effect of the confidence interval within that state.

In addition, certain conditions used by states to determine AYP status were not evaluated as part of this study. Some schools identified in our illustration as failing to make AYP would make it because they met their state's safe harbor provisions. Some would now also pass under the growth-model pilot underway in a handful of states, such as Ohio. In this respect, our findings do underestimate the actual AYP performance of some of the schools in the sample. Conversely, a few schools identified as making AYP might actually fail to make it because they did not meet their state's average daily attendance requirement or because they did not test 95% of a particular subgroup(s) within their student population. While we concede that our results may understate actual AYP performance in some cases, we believe the study provides a relatively accurate and useful prediction of how schools generally fare under the *base* AYP rules. That is, if NCLB was intended to get 100% of students, including those within subgroups, across the proficiency bar, the study illustrates how well the sample schools fared relative to this goal and its benchmarks.

With these limitations considered, we believe this study illuminates the inconsistency of AMOs and proficiency cut scores and other rules for determining AYP status across states. It does not, however, necessarily replicate with precision the performance and AYP status of the sample schools within their own state, or predict with complete consistency their status if students took the exams required by other states.

[6] This means that students are offered questions at a level of difficulty that reflect their current performance rather than their current grade. For example, a high-performing third-grader might receive questions at the fifth-grade level, while her lower-performing peer might receive questions pegged at the first-grade level.